Министерство общего и профессионального образования Российской Федерации

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ГЕОДЕЗИИ И КАРТОГРАФИИ

Людмила Николаевна Чабан

ТЕОРИЯ И АЛГОРИТМЫ РАСПОЗНАВАНИЯ ОБРАЗОВ.

Учебное пособие.

Москва
2004

УДК
ББК

Рецензент:
Зав. кафедрой ВТиАОИ МИИГАиК, д.т.н. И.Г.Журкин

Л.Н.Чабан. Теория и алгоритмы распознавания образов. Учебное пособие.
М.: МИИГАиК. 2004. – 70с.

© Московский государственный университет геодезии
и картографии (МИИГАиК)

Введение.

Современные информационные системы и технологии включают в себя большое количество процедур, моделирующих или поддерживающих процесс интеллектуального анализа данных. К простейшим процедурам такого типа относится любая классификация количественных данных по заданным пользователям критериям, более сложные обеспечивают анализ сцен, процессов, явлений с целью выделения объектов с заданными характеристиками или свойствами.

Процедуры подобного типа присутствуют не только в задачах анализа аэрокосмических изображений, но и при обработке сигналов в технических системах, в медицинской диагностике, биологии, социологии, банковском деле и других областях человеческой деятельности. По мере расширения сферы применения геоинформационных технологий и усложнения процедур геоинформационного моделирования процедуры анализа и классификации совокупностей данных, объектов и структур занимают все более значимое место и в геоинформационных системах нового поколения.

В системах обработки и анализа данных дистанционного зондирования (ДЗ) спектр процедур подобного типа представлен наиболее широко. Практически весь процесс тематического дешифрирования аэрокосмической информации состоит из поэтапной группировки и дальнейшего преобразования данных с целью создания совершенно определенной, проблемно-ориентированной картины земной поверхности. Значительная часть этих этапов обеспечивается методами и алгоритмами, входящими в специализированные инструментальные пакеты, и задача обработчика состоит в создании наиболее эффективной схемы классификации данных применительно к поставленной прикладной задаче.

Проектирование любой системы тематического анализа и классификации информационных объектов и структур, независимо от сферы ее применения, требует от специалиста знания не только общих принципов проектирования информационных систем, но также основных методов и алгоритмов, обеспечивающих этот процесс, их рабочих характеристик и возможностей применения для анализа тех или иных типов информации.

Исследованием и разработкой методов, алгоритмов и систем для решения таких задач на ЭВМ занимается дисциплина, получившая название **распознавание образов.**

Настоящее пособие представляет собой конспект первой части двухсеместрового лекционного курса, читаемого на специальности «Информационные системы в геодезии и картографии», и примерно соответствует содержанию односеместрового курса на специальности «Исследование природных ресурсов» Московского государственного

университета геодезии и картографии (МИИГАиК). Здесь рассматриваются основные концепции теории распознавания образов, а также методы и алгоритмы распознавания, применяемые при классификации многомерных количественных данных. Большая часть этих методов возникла практически одновременно с появлением ЭВМ и в настоящее время составляет «классическую» часть любого курса по распознаванию образов. Реализующие их алгоритмы присутствуют в той или иной конфигурации во всех пакетах обработки данных ДЗ и используются преимущественно при обработке многозональных или многослойных растровых изображений земной поверхности. Безусловно, сфера применения рассматриваемых подходов не ограничивается указанными приложениями. Тем не менее, при рассмотрении данной группы методов распознавания наибольшее внимание уделено именно этой сфере применения, поскольку материалы ДЗ широко используются при составлении и обновлении карт различного тематического содержания.

В связи с тем, что курс рассчитан в основном на будущих разработчиков прикладных геоинформационных технологий, значительное место в пособии отведено вопросам построения систем распознавания, особенностям использования рассматриваемых методов и алгоритмов в прикладных задачах и специфике их реализации в известных пакетах обработки данных ДЗ.

Последний раздел данного пособия носит чисто ознакомительный характер, так как представляет интерес в основном для специалистов, непосредственно занимающихся разработкой и эксплуатацией систем распознавания. Но некоторые сведения из этого раздела могут быть полезны при изучении современных направлений в методологии распознавания образов. Более подробно с материалом этого раздела можно ознакомиться в монографиях [1,20,31].

При составлении пособия основными литературными источниками были известные монографии по распознаванию образов [1, 2, 3] и, частично, учебники по физическим основам и обработке данных дистанционного зондирования [4,11]. В разделе 6 использовалась также классическая монография по кластерному анализу [6] и некоторые обобщения из [7-9,24]. Изложение раздела 7 основывается на соответствующих разделах монографии [5], где данные вопросы изложены в наиболее общем виде. В изложении некоторых прикладных аспектов использовалось «Руководство пользователя» по пакету ERDAS Imagine, на котором в университете проводятся практические занятия по обработке и анализу данных ДЗ, а также личный практический опыт автора [13-16]. Ссылки на дополнительную литературу по отдельным вопросам для углубленного изучения предмета приведены в тексте пособия.

1. Основные понятия теории распознавания образов.

Распознавание образов, как методология принятия решений на основе результатов наблюдений за объектами и процессами окружающего мира, возникло значительно раньше современных компьютерных систем и технологий. Первые методы распознавания разрабатывались для электронных аналоговых систем и рассматривались в рамках теории обработки сигналов. В процессе развития вычислительной техники и информационных технологий эта дисциплина, как и само понятие образа, претерпела значительные изменения и продолжает интенсивно развиваться. Сейчас трудно строго определить класс задач, который относится исключительно к распознаванию образов, равно как и дать строгое определение самого предмета изучения.

В обыденном понимании образ включает целую совокупность наших индивидуальных ощущений, представлений и умозаключений. Распознавание образов - повседневная неотъемлемая составляющая деятельности человеческого мозга. Поэтому в спектре компьютерных дисциплин задачи распознавания относятся к проблематике **искусственного интеллекта.**

В самом общем случае в качестве образа может рассматриваться любая информационная модель объекта или процесса абстрактного или реального мира. Отличительная особенность такой модели в задаче распознавания - это использование только того подмножества характеристик объектов исследования, которое обеспечивает выделение одной или нескольких групп объектов совершенно определенного типа. Целью процедуры распознавания является ответ на вопрос: относится ли объект, описанный заданными характеристиками, к интересующим нас категориям и если относится, то к какой именно.

Исходя из этого, можно сказать, что **образ** - это описание объекта или процесса, позволяющее выделять его из окружающей среды и группировать с другими объектами или процессами для принятия необходимых решений.

Те категории объектов, которые мы хотим выделить или на которые хотим разделить все множество образов в процессе распознавания, обычно называют **классами**. Понятие класса в распознавании появилось существенно раньше, чем в объектных информационных моделях. И способ задания классов здесь определяется не столько сущностью предмета исследования, сколько особенностями имеющейся информации об объектах и способами ее представления.

Для системы обработки информации **образ -** это совокупность данных об объекте или явлении, включающая параметры и связи. Параметры представляют собой количественные характеристики, полученные с помощью измерительных систем или математических моделей. Связи могут

описывать как внутреннюю структуру образа, так и особенности его поведения, если мы имеем дело с динамическим объектом или процессом.

Любой алгоритм распознавания можно представить как абстрактную функциональную систему R, состоящую из трех компонент:

$$R=\{A,S,P\}, \qquad (1.1)$$

где

$A=\{A_k\}$, $k=1,...,K$ – алфавит классов – множество категорий, по которым мы должны распределить наши образы,

$S=\{S_j\}$, $j=1,...,n$ - словарь признаков - множество характеристик, из которых составляется описание образа,

$P=\{P_l\}$, $l=1,...,L$ - множество правил принятия решения.

Функционирование этой системы сводится к следующему: на вход подается образ – некоторая конфигурация из элементов множества S, к ней применяется определенная последовательность правил из P, в результате конфигурации присваевается индекс, соответствующий одному из элементов множества A. Качество функционирования системы определяется тем, насколько часто присвоенный образу индекс совпадает с ожидаемым нами результатом.

Компоненты A,S представляют собой информационную часть системы, а P – методологическую. Ясно, что смысл понятия **класс** для различных способов описания образов будет различным. В свою очередь, способ описания образа зависит от физической природы объектов распознавания и возможностей формализации соответствующих им понятий. Методы принятия решений, естественно, взаимосвязаны со способом представления объектов распознавания. Поэтому любая система распознавания включает и процесс **синтеза образов**, то есть формирования описаний объектов распознавания и их классов, и **анализа образов**, то есть сам процесс принятия решений.

В зависимости от особенностей информационных компонент системы R, выделяют три подхода к задаче распознавания образов: 1) принцип сравнения с эталоном; 2) принцип кластеризации; 3) принцип общности свойств.

Принцип сравнения с эталоном применяется в тех случаях, когда каждому классу A_k можно сопоставить конечный набор эталонных образов $\Omega_k=\{\omega_m,\ m=1,...,M_k\}$. Поэтому принцип сравнения с эталоном иначе называют принципом перечисления. В этом случае процесс распознавания заключается в простом сопоставлении образов, поступающих на вход распознающего устройства или алгоритма, с эталонами Ω_k классов A_k , на основе выбранной меры сходства.

Принцип сравнения с эталоном - один из первых подходов, возникших при построении технических систем распознавания, когда возможности

вычислительных устройств были весьма ограничены. Тем не менее, он и сейчас применяется, в частности, в аналоговых и аналогово-цифровых системах распознавания. Распространенным примером применения такого подхода является распознавание печатных шрифтов. В тематической обработке изображений он может применяться при распознавании текстур и выделении объектов определенной формы.

Принцип кластеризации. Если признаки представляют собой наборы измерений (параметров) без каких-либо явным образом заданных взаимосвязей, то образ можно представить как n-мерный вектор в признаковом пространстве **X**:

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Каждому классу A_k сопоставляется некоторое множество векторов в пространстве **X**. В результате признаковое пространство разбивается на области, соответствующие классам A_k, которые называют кластерами или таксонами. Одному классу может соответствовать любое количество областей пространства **X**. Эти области могут перекрываться, но методология принятия решения должна в конечном итоге обеспечивать однозначное отнесение точки пространства к определенному классу.

Принцип кластеризации (таксономии) широко применяется в различных прикладных направлениях при обработке количественных данных, в том числе и в системах компьютерного анализа многозональных и спектрозональных аэрокосмических изображений (классификация по спектральным признакам).

Принцип общности свойств существенно использует связи между элементами образа. Как правило, он применяется в тех случаях, когда множество образов каждого класса слишком велико, чтобы получить надежное описание конечного числа эталонов, но можно выявить достаточное количество отличительных особенностей классов по конечным выборкам образов. Выявленные свойства кодируются на основе подходящей модели и хранятся в памяти в виде некоторых структур, функций или отношений. В процессе распознавания производится анализ образа по схеме, позволяющей выявить необходимые свойства образа; затем они сопоставляются со свойствами классов A_k. Обобщающим свойством может быть сам алгоритм порождения образов; в этом случае классы образов задаются алгоритмами порождения структур определенного вида.

Перечисленные подходы к распознаванию относятся в основном к процессу **синтеза образов**, то есть к способу представления информации - компонент A и S системы R. Методы **анализа образов**, то есть правила принятия решений, в большинстве случаев взаимосвязаны со способом

синтеза образов, хотя довольно часто тот или иной аппарат принятия решений, например, статистический, может применяться при различных способах синтеза.

В методологии принятия решений при распознавании образов также выделяют три основных направления: 1) эвристические методы; 2) математические методы; 3) лингвистические (синтаксические) методы.

Эвристические методы основываются на опыте и интуиции разработчика системы распознавания. Как правило, эти методы ориентированы на решение конкретного типа задач распознавания и непосредственно привязаны к способу синтеза образов. Наиболее часто применяются при использовании принципа сравнения с эталоном и принципа общности свойств.

Математические методы опираются на использование классического математического аппарата: методов линейного программирования, корреляционного анализа, теории статистических решений и т.п. Применяются в тех случаях, когда признаки представлены измерениями (параметрами), а их связи могут быть описаны в виде аналитических зависимостей. В той или иной степени математические методы используются во всех трех подходах к распознаванию.

Математические методы обычно подразделяют на детерминистские и статистические. В задачах анализа данных ДЗ наиболее часто применяют статистические методы. Это обусловлено высокой степенью изменчивости условий наблюдения и самих объектов исследования, а также потерей информации на разных этапах проведения ДЗ.

Лингвистические (синтаксические, структурные) методы применяются в тех случаях, когда образ представляет собой некоторую структуру, состоящую из так называемых непроизводных (первичных) элементов и признаков, описывающих связи между ними. В этих методах широко используется аппарат алгебры логики и теории формальных языков. При обработке данных ДЗ и в геоинформационных технологиях эти методы могут использоваться при анализе пространственных структур, некоторые из них применяются также в экспертных системах.

Современный уровень развития вычислительных средств позволяет комбинировать в системах распознавания как различные подходы к описанию образов, так и методы, использующиеся в процессе распознавания. Рациональное использование этих подходов является главной залачей разработчика системы распознавания.

2. Принципы построения систем распознавания образов.

В отличие от рассмотренной выше абстрактной системы, реальная **система распознавания** – это функционально взаимосвязанная совокупность методов и технических средств, осуществляющая процесс синтеза и анализа образов.

Прежде чем приступать к построению системы распознавания, необходимо проанализировать всю доступную информацию об объектах исследования и решить следующие вопросы.

1. Какими общими характеристиками и свойствами обладают объекты исследования и чем они различаются.
2. Если необходимые характеристики могут быть получены в результате измерений, какова точность этих измерений.
3. Существует ли подходящая модель (модели) для формального описания и анализа данных характеристик.

На основании проведенных исследований определяется тип и структура системы распознавания.

По физической природе характеристик-признаков образов системы распознавания подразделяются на **простые и сложные**. Примером простой системы распознавания в задачах обработки данных ДЗ является классификация пикселов многозональных сканерных или оцифрованных спектрозональных изображений. В этом случае признаками являются спектральные отражательные свойства объектов земной поверхности. Как только в процесс классификации привлекаются другие типы данных, система становится сложной.

Сложные системы распознавания могут быть **одноуровневыми и многоуровневыми.** В одноуровневых системах распознавание осуществляется на основе одного словаря признаков одним алгоритмом распознавания. В многоуровневых системах результаты распознавания, полученные на одном этапе, используются в качестве исходных данных на следующем.

Большая часть технологий тематической обработки данных ДЗ при решении прикладных задач реализуется сложными многоуровневыми системами распознавания. Разбиение схемы решения задачи на уровни называется **декомпозицией.** Способ декомпозиции системы распознавания обычно связан с моментами включения в процесс распознавания специалиста-аналитика данных, то есть с **интерактивным режимом** обработки.

Необходимость в интерактивном режиме чаще всего возникает на этапе синтеза образов из-за неполноты исходной информации или трудностей ее формального описания. В многоуровневых системах важной функцией

аналитика данных является оценка качества распознавания на текущем этапе и возникающей из-за этого потери информации на следующем.

По полноте исходной информации системы распознавания подразделяются на системы **с обучением, без обучения и самообучающиеся системы.**

В **системах без обучения** предполагается, что имеющаяся информация и выбранный принцип распознавания позволяют безошибочно разделить все необходимые классы. Типичным примером систем без обучения являются системы, основанные на принципе сравнения с эталоном (принципе перечисления). Системы без обучения могут быть построены и на принципе кластеризации, когда векторы-образы каждого класса образуют неперекрывающиеся компактные группы.

Если имеющихся данных об объектах исследования недостаточно для их точного разделения, распознавание без обучения приводит к большим ошибкам. И наоборот, количество информации об объектах может быть настолько большим, что для выделения нужных классов нет необходимости использовать ее в полном объеме. В этих случаях применяют **системы распознавания с обучением.**

Системы с обучением широко используются при тематической обработке данных ДЗ. Процедура обучения обычно выполняется специалистом-аналитиком данных в интерактивном режиме. Следует отметить, что любые процедуры классификации с обучением более трудоемки и качество распознавания иногда больше зависит от личного опыта аналитика данных и понимания им сути поставленной задачи, чем от эффективности самого алгоритма распознавания.

Чаще всего системы с обучением основываются так или иначе на принципе общности свойств. При этом образы не обязательно должны быть представлены какой-то структурой. Это может быть просто множество параметров-измерений, подчиняющихся в каждом классе определенному закону статистического распределения, или множество каких-то атомарных (неделимых) элементов, присутствующих в каждом классе в определенной комбинации. Такие множества в обобщенной теории образов называют свободной конфигурацией [7,8]. Задача обучения заключается в выявлении типичных для каждого класса свойств образов или взаимосвязей между ними. Обучение в системах, использующих принцип кластеризации, заключается в поиске такого разбиения всего множества вектор-образов на кластеры, которое обеспечивало бы минимальные ошибки в процессе распознавания.

В **самообучающихся системах** процедура обучения выполняется в самом процессе распознавания специальным алгоритмом. Для оценки качества распознавания используется некоторый функционал, связанный с ошибками распознавания, который в процессе обучения максимизируется

или минимизируется. Одним из широко известных типов такого функционала является сумма квадратов ошибок (НСКО- критерии)*.

Рис. 2.1. Общая схема построения системы распознавания.

* В некоторых современных работах [19,31] этот функционал, связанный со средней величиной ошибки классификации, называют **функцией потерь**, обобщая таким образом понятие, возникшее из игровой постановки задачи распознавания.

Большое значение при декомпозиции сложных систем имеют требования к **точности распознавания**, то есть к допустимой вероятности ошибок. Если в одном алгоритме распознавания используются признаки различной физической природы или полученные из принципиально различных моделей, то величина ошибки может оказаться непредсказуемой. Поэтому декомпозиция системы должна осуществляться так, чтобы можно было надежно оценить точность распознавания на отдельных этапах решения задачи и скорректировать, при необходимости, всю схему решения.

После декомпозиции сложной системы каждый уровень можно рассматривать как отдельную подсистему распознавания R со своим списком классов, словарем признаков и правилами классификации. Общая схема построения системы распознавания показана на рис. 2.1.

Оценка эффективности системы распознавания выполняется для всей системы в целом, поскольку она может существенно зависеть от способа декомпозиции задачи. Ее можно оценивать по следующим показателям.

1. Итоговая точность распознавания.
2. Общее время распознавания.
3. Стоимость разработки и эксплуатации системы.

При построении интерактивных систем распознавания, в частности, обработки данных ДЗ, в оценку эффективности системы могут быть включены затраты на подготовку специалистов-аналитиков, участвующих в процессе распознавания, поскольку от их квалификации часто зависит итоговая точность и общее время решения задачи.

3. Особенности систем и методов распознавания в задачах тематической обработки данных ДЗ.

Система компьютерной тематической обработки данных ДЗ в большинстве случаев является сложной многоуровневой системой распознавания с большим количеством интерактивных процедур.

Основным типом входных данных в такой системе являются цифровые изображения земной поверхности, полученные с различных типов съемочной аппаратуры. Это могут быть

- многозональные изображения, полученные с космических и самолетных сканирующих систем;
- панхроматические сканерные изображения;
- оцифрованные спектрально совмещенные многозональные космические и аэрофотоснимки;
- оцифрованные спектрозональные космические и аэрофотоснимки;
- оцифрованные панхроматические аэрокосмические снимки;
- радиолокационные изображения.

Значение яркости каждого элемента (пиксела) серого полутонового изображения чаще всего представляется в шкале 0-255 (1 байт информации на пиксел), которую называют **динамическим диапазоном**. Многозональные изображения представляются как набор совмещенных по каналам серых полутоновых изображений. Спектрозональные изображения обычно раскладываются на три серых полутоновых изображения по цветовым компонентам, наиболее часто в системах RGB (красный-зеленый-синий) или HSI (чистый цвет-насыщенность-яркость).

Кроме этих данных, при тематической обработке обычно используются
- топографические карты;
- тематические карты;
- данные наземных и аэровизуальных обследований;
- фондовые материалы (экономико-географические данные по территории, справочные данные по тематике задачи, данные предыдущих обследований и т.п.).

Таким образом, **первая особенность** систем тематической обработки данных ДЗ как систем распознавания - комплексное использование данных из принципиально различных источников. С одной стороны – это количественные характеристики, полученные с помощью измерительных систем (аэрокосмических аппаратов, приборов для выполнения наземных обследований), с другой стороны - описательные признаки изображенного ландшафта, применяющиеся при визуально-инструментальном дешифрировании снимков и часто трудно формализуемые. Из-за сложности и многообразия данных в системах тематической обработки данных ДЗ важное место отводится оператору-аналитику и организации интерактивного режима.

Вторая особенность - наличие этапа предварительной обработки данных, предшествующего процессу распознавания. При этом характер предварительной обработки (процедур коррекции) непосредственно взаимосвязан с методами распознавания, которые применяются для решения задачи.

Третья особенность рассматриваемой системы состоит в том, что в ней одновременно или попеременно используются две взаимосвязанные системы координат: 1) система координат изображения; 2) система координат измеряемых количественных параметров. Учет взаимосвязи этих двух систем координат часто оказывается наиболее сложной задачей при построении процедур распознавания, особенно для оценивания ошибок.

При оценке качества результата обработки в системе тематической обработки данных ДЗ обычно используют два понятия: **точность и достоверность. Точность** связана с системой координат изображения и теми характеристиками объектов земной поверхности, которые рассчитываются в этой системе координат (положение, расстояния, площади и т.п.).

Достоверность - с правильной тематической интерпретацией объектов выделенных классов. Ясно, что если объект неправильно интерпретирован, это отразится на точности последующих количественных (линейных и площадных) оценок по исследуемым тематическим категориям в системе координат изображения. Поскольку тематическая интерпретация выделенных объектов земной поверхности является конечной целью данной задачи, процент неправильной интерпретации - это итоговая ошибка распознавания.

Рассмотрим, как формируется эта итоговая ошибка.

Прежде всего, она связана со способом **декомпозиции** системы распознавания. Существуют два основных подхода к декомпозиции. **Первый подход** состоит в первоначальной классификации всех точек изображения в системе координат измеренных параметров-характеристик (спектральных яркостей, оптических плотностей), а затем объединении классифицированных точек в пространственно связные объекты уже с системе координат изображения с использованием интерактивного пространственного анализа полученных результатов. **Второй подход,** традиционно применяющийся в интерактивном визуальном дешифрировании - это предварительное выделение самих объектов, описание их интегральных яркостных и других характеристик, а затем их классификация поэтим признакам.

Первый подход привлекателен тем, что при достаточном для разделения исследуемых классов количестве параметров-признаков он позволяет получить наиболее высокую точность границ объектов в системе координат изображения. Более того, эти границы будут наиболее объективны и оценить вероятность ошибок здесь существенно легче.

Если измеряемых параметров недостаточно для разделения тематических классов, обычно применяется второй подход: сначала выделяются границы объектов, затем формируется система признаков, включающая одновременно и характеристики отдельных пикселов, и их пространственные взаимосвязи. Недостатком этого подхода при компьютерной реализации является сложность выбора математической модели, адекватно описывающей исследуемые объекты в системе координат изображения. Структурные и особенно текстурные признаки объектов земной поверхности недостаточно устойчивы как к изменению пространственного разрешения, так и к условиям освещенности[*]. Поэтому алгоритмы такого

[*] Спектральная яркость пикселов, относящихся к одному классу объектов подстилающей поверхности, также зависит от условий освещенности, особенно на изображениях высокого разрешения. Тем не менее, существуют специальные приемы для решения этой проблемы, рассмотренные, например, в работах [22,25,27].

типа далеко не универсальны и обычно разрабатываются для решения узко специальных задач с использованием конкретного типа информации.

Первый подход обычно используется при обработке многозональных изображений, второй - крупномасштабных панхроматических снимков и радиолокационных изображений.

В принципе, поэлементная классификация изображений с обучением близка ко второму подходу, поэтому при использовании таких процедур основная нагрузка приходится на организацию процесса обучения. Здесь необходимо учитывать ряд особенностей формирования признаков, которые будут рассмотрены в соответствующем разделе.

Общей важной характеристикой цифровых изображений всех перечисленных типов, влияющей на величину итоговых ошибок, является **размер элемента разрешения (пиксела) на земной поверхности**. В соответствии с физическими основами ДЗ, яркость (или цвет) каждого пиксела изображения характеризует спектральные отражательные свойства соответствующего ему участка земной поверхности. В зависимости от разрешения, в пределах одного пиксела могут интегрироваться разнообразные элементы ландшафта с различными спектральными отражательными свойствами. Поэтому при распознавании только по яркостным признакам, даже если точность выделения границ объектов в системе координат изображения высока, неизбежны ошибки интерпретации отдельных пикселов и даже мелких объектов. Для сокращения ошибок лучше задавать избыточное количество классов, а потом группировать их с учетом структурных признаков в системе координат изображения, а также с использованием картографических и справочных данных.

В настоящее время много внимания уделяется методам разделения смесей классов в пределах одного пиксела [25, 26]. Однако для решения этой задачи необходимо сначала создать банк данных «чистых» классов, то есть надежно разделяющихся на данном изображении по спектральным характеристикам.

При анализе сканерных изображений размер элемента разрешения фиксирован, поэтому является жестким ограничением для задач распознавания. Для фотоснимков разрешение зависит от режима оцифровки, ограничением является размер зерна конкретного типа пленки. Но необходимо помнить, что даже небольшое изменение размера элемента разрешения может привести к изменению рисунка изображения, а следовательно, к изменению дешифровочных признаков. Поэтому при распознавании образов в системе координат изображения (по структурным или текстурным признакам) необходимо обязательно исследовать зависимость признаков от этого фактора.

Другим внешним по отношению к системе обработки источником ошибок являются **погрешности измерений**, возникающие в процессе проведения ДЗ, - геометрические и яркостные искажения. И если учет и устранение первого типа искажений является чисто технической задачей, то влияние искажений второго типа обычно полностью устранить невозможно. Чаще всего для яркостных характеристик изображения выполняется радиационная (атмосферная) и радиометрическая (приборная) коррекция. Вариации яркостных характеристик объектов земной поверхности, вызванные особенностями текущих природно-климатических условий (например, влиянием ветра или предшествующих ДЗ осадков), а также индикатрисы отражения, особенно для высокоширотных территорий, приходится учитывать тем или иным способом в процессе распознавания, поскольку влияние этих факторов на различные классы объектов земной поверхности не одинаково.

4. Основные принципы построения систем распознавания данных ДЗ.

Общая схема системы распознавания данных ДЗ для решения прикладных задач состоит из блоков, представленных на рис. 4.1.

Рассмотрим наиболее важные особенности каждого из этапов.

1. **Нормализация** включает процедуры, устраняющие основные типы искажений исходных данных. На этом этапе следует использовать только те процедуры, которые не приводят к значительным потерям информации: еще неизвестно, какие характеристики изображения наиболее существенны для решения задачи, а какими можно пренебречь. Нужно помнить, что геометрические преобразования могут привести к искажению исходных яркостных характеристик объектов подстилающей поверхности. На этом этапе их следует выполнять только в тех случаях, когда имеются наземные измерения, требующие строгой координатной привязки тестовых участков к изображению. Тем более нельзя трансформировать изображение в какую-то картографическую проекцию, если предполагается выполнение его классификации по спектральным признакам.

2. **Визуализация и комплексный анализ данных**. На этом этапе производится анализ информативности имеющихся данных с точки зрения поставленной задачи и выбор принципа декомпозиции системы распознавания. При наличии нескольких изображений на обследуемую территорию выполняется их пространственное совмещение (приведение к единой системе координат). Для визуального анализа обычно применяются процедуры поднятия контраста (растягивание ненулевого интервала значений яркости на весь динамический диапазон). При анализе многозональных изображений применяется RGB-синтез в удобной для анализа цветовой палитре, а также **анализ главных компонент**, который мы рассмотрим в

соответствующем разделе. Описание основных методов улучшения качества изображений можно найти в [10,11,19,23], а также в разделе «Enhancement» руководства пользователя по пакету ERDAS Imagine.

Процедуры улучшения визуального восприятия могут, однако, отрицательно повлиять на качество распознавания, если они хоть каким-то образом связаны с пересчетом значений яркости на исходном изображении. Выполняя подобные преобразования цифрового изображения, всегда необходимо проверять, не изменяет ли оно те свойства объектов исследования, которые будут использоваться в системе распознавания.

3. На основе комплексного анализа данных формируется **общая схема решения задачи**. Схема включает основной набор процедур классификации, последовательность их применения, перечень входных и выходных данных и вспомогательных процедур их обработки и анализа. Как уже говорилось выше, выбор таких процедур обработки изображений должен быть согласован с принципами классификации и не должен приводить к искажению тех характеристик изображения, которые могут быть использованы как признаки в системе распознавания.

4. **Выделение признаков** представляет собой интерактивный процесс, включающий ряд автоматических процедур преобразования видеоданных. В системах распознавания, основанных на классификации пикселов изображения, очень важно правильно выбрать количество классов в системе координат признаков. Если выбранное количество классов значительно меньше количества различающихся по этим измерениям объектов земной поверхности, ошибки классификации могут оказаться недопустимо высокими. Для выбора необходимого количества исходных классов в таких системах применяется **неконтролируемая классификация** и **сегментация изображений** [4,13,14,23]. При большом количестве измерений-признаков для сокращения размерности задачи применяется **корреляционный и факторный анализ данных.**

При распознавании по интегральным признакам в системе координат изображения часто важно усиление определенных структурных и текстурных признаков. В этих случаях в качестве вспомогательных обычно используются процедуры **низкочастотной и высокочастотной фильтрации** и различные операторы для подчеркивания геоиндикационных признаков [10,12]. При использовании структурных признаков объектов необходимо предварительное устранение геометрических искажений.

5. **Классификация** - это основной блок распознавания для решаемой задачи, хотя отдельные процедуры распознавания могут присутствовать и на этапе выделения признаков, и на этапе постклассификационной обработки. Но именно результат, полученный на этапе классификации изображения, является основой для интерпретации и последующих процедур перехода от

изображения к тематической карте. Процесс классификации может осуществляться в несколько этапов или в несколько итераций, в зависимости от выбранной схемы распознавания. В ряде случаев результаты, полученные на этом этапе, требуют пересмотра всей схемы обработки в целом.

6. **Тематическая интерпретация** заключается в сопоставлении каждому из выделенных классов некоторой тематической категории определенной совокупности объектов земной поверхности. Здесь оцениваются реальные ошибки классификации, то есть достоверность результата. При неудачном выборе количества или описаний исходных классов может быть принято решение о повторной классификации на основе другого набора классов и системы признаков. При удовлетворительном результате выбирается схема постклассификационной обработки для перехода к тематической карте.

7. **Постклассификационная обработка** представляет собой набор интерактивных и автоматических процедур, необходимых для формирования карты объектов исследования, расклассифицированных в соответствии с тематикой задачи. К таким процедурам относятся группировка и редактирование выделенных классов, устранение обнаруженных ошибок классификации. В процессе постклассификационной обработки также могут быть использованы процедуры распознавания, в том числе структурные.

8. **Подсчет количественных характеристик** включает площадные оценки по выделенным классам объектов, а также точечные оценки в соответствии с физическими моделями исследуемых процессов и явлений, оценки динамики процессов при обработке мультивременных данных и т.п. Этот этап, не имеющий прямого отношения к процессу распознавания, тем не менее, важен для его оптимизации. Накопление и анализ характеристик выделяемых классов объектов позволяет совершенствовать как саму систему распознавания, так и отдельные ее этапы.

9. Процесс **получения тематических карт** на основе результата обработки включает процедуры преобразования изображения к нужной картографической проекции, интерактивное редактирование результата в соответствии с точностными требованиями к картографическим материалам заданного масштаба, а также растрово-векторное преобразование при использовании векторных ГИС. Выполнение растрово-векторного преобразования требует, как правило,сглаживания результата в пределах некоторых участков для получения однородных по классам контуров. Именно с точки зрения этого этапа наиболее эффективны такие процедуры классификации, которые обеспечивают наибольшую однородность изображения по классам и наиболее четкие границы между объектами определенных тематических категорий. Если результат классификации не удовлетворяет этим условиям, процедуры сглаживания могут привести к искажению реальных границ между классами.

Рис. 4.1. Общая схема тематической обработки данных ДЗ.

В настоящее время существует большое количество специализированных пакетов, содержащих набор необходимых автоматических процедур для построения систем тематической обработки данных ДЗ. Наиболее известные профессиональные продукты - ERDAS Imagine, ER Mapper (Австралия), EASY/PAGE (PCI) (Канада), ILVIS GIS (Голландия). Наиболее распространенными на отечественном рынке программных продуктов являются пакеты ERDAS Imagine и ER Mapper, а

также небольшие настольные системы для PC – IDRISI, EPPLE7, GRASS. С начала 90-х все большее распространение получает пакет ENVI (США), включающий некоторые новые средства анализа данных ДЗ и позволяющий формировать гибкие схемы обработки благодаря наличию внутреннего объектного языка.

Все такие пакеты включают стандартный набор классификаторов многозональных (многослойных) растровых данных и в этом аспекте различаются в основном интерфейсом и сервисными процедурами интерактивного анализа. Несколько отличается также способ реализации отдельных алгоритмов. С точки зрения построения прикладной системы распознавания наибольший интерес представляют два инструментальных аспекта этих пакетов: 1) функции, обеспечивающие предварительный анализ данных и обучение классификаторов; 2) функции, обеспечивающие тематическую интерпретацию результатов классификации и их редактирование.

В первом аспекте наибольшее значение имеет способ визуализации многозональных изображений, то есть выбор линейной комбинации каналов и цветовой палитры для результата синтеза [32]. Неудачный выбор способа синтеза каналов приводит к потере информации при визуальном анализе и, как следствие, ограничивает возможности выбора необходимых классов и описания их признаков. Поэтому при предварительном визуальном анализе изображения часто бывает необходимо апробировать достаточно большое количество различных режимов визуализации.

Следующим важным фактором является набор процедур, обеспечивающих процесс описания классов в заданной системе признаков. При классификации пикселов многозонального изображения большое значение имеет определение наиболее точных границ между классами в признаковом пространстве, то есть в пространстве спектральных яркостей. Возможность формирования этих границ только путем выделения на изображении эталонных участков изначально ограничивает точность классификации любым алгоритмом, так как границы выделяемых тематических категорий в пространстве признаков далеко не всегда можно точно аппроксимировать известными аналитическими функциями, которые обыкновенно используются в стандартных классификаторах. Более того, некоторым классам может соответствовать несколько не связных ограниченных областей признакового пространства. В этом случае большой интерес представляет анализ положения классов непосредственно в пространстве спектральных признаков и возможность предварительной неконтролируемой классификации (кластеризации) в различных режимах с последующим анализом спектральных сигнатур полученных кластеров.

Для проведения исследований с целью построения эффективной схемы классификации наиболее удобны средства пакета ERDAS Imagine. Растровый

редактор (Viewer) обеспечивает большое количество различных средств интерактивного визуального анализа и способов формирования обучающих данных. В пакете имеется достаточно большое количество средств настройки классификаторов, анализа обучающих выборок и оценки качества классификации.

Для знакомства с основными процедурами классификации изображений и построения небольших, но гибких схем тематической обработки можно использовать пакет IDRISI. В этом пакете имеется достаточно широкий набор средств тематического анализа и классификации, хотя возможности настройки самих алгоритмов несколько ограничены в сравнении с ERDAS Imagine. То же самое можно сказать и о классификационном блоке пакета ER Mapper.

Контрольные вопросы к разделам 1-4.

1. Перечислите основные методологические подходы к задаче распознавания образов. Какие методологические подходы могут применяться при решении задач: 1) распознавания картографических условных знаков; 2) классификации административных единиц одновременно по площади территории и численности населения; 3) распознавания сельских домовладений на снимке или плане как определенных конфигураций объектов правильной геометрической формы?

2. Что такое сложная система распознавания? По каким критериям может осуществляться декомпозиция сложной системы?

3. Каковы достоинства и недостатки систем распознавания с обучением и без обучения? Что представляет собой самообучающаяся система?

4. По каким критериям оценивается эффективность системы (алгоритма) распознавания? Какими путями можно повысить эффективность системы?

5. Приведите примеры задач распознавания, которые могут быть решены: 1) системами распознавания без обучения; 2) только системами с обучением.

6. Каковы два основных подхода к построению систем распознавания при тематической обработке данных ДЗ? Какой из подходов обеспечивает более высокую точность площадных оценок исследуемых характеристик объектов земной поверхности и почему?

7. Какие факторы влияют на качество распознавания в системах тематической обработки аэрокосмических изображений: 1) по спектральным (яркостным, тоновым) признакам; 2) по текстурным (структурным) принакам? Какие виды нормализации изображений наиболее важны в каждом из этих случаев?

5. Задача классификации образов, представленных набором измеряемых параметров. Геометрическая и аналитическая интерпретация.

Спектр задач классификации образов, представленных наборами измеряемых или вычисляемых параметров-характеристик, достаточно обширен. К наиболее известным примерам использования таких классификаторов в геоинформационных технологиях можно отнести тематическую группировку объектов по количественным значениям атрибутов при картографической генерализации, при создании легенд к картам, описывающим динамические процессы на земной поверхности, и в целом ряде задач прикладного геоинформационного моделирования. При тематической обработке материалов дистанционного зондирования к этому типу задач распознавания относятся практически все процедуры классификации многозональных или многослойных растровых изображений. Если для каждой точки цифрового изображения имеется набор данных, представленный действительными или целыми числами - результатами измерений n параметров, эту точку можно рассматривать как образ-вектор

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

в n-мерном пространстве X значений этих параметров.

Такие задачи встречаются не только при классификации точек изображения и при группировке объектов по значениям атрибутов в геоинформационных технологиях, но и при анализе количественных данных в экономике, демографии, биологии и многих других предметных областях.

Как уже упоминалось в первом разделе, методы и алгоритмы, реализующие данный подход, основываются на **принципе кластеризации,** то есть сопоставлении каждому классу одного или нескольких интервалов значений каждого признака. В данной постановке задачи параметр x_j является j-й координатой пространства **X**, j=1,…,n, которое называют **признаковым пространством**. При конкретном перечне (алфавите) классов $\{A_k\}$, k=1,...,K, значения признаков для образов различных классов могут совпадать или перекрываться по одним координатам и различаться по другим. Если выбрать такой набор признаков, чтобы образы разных классов различались хотя бы по одной координате, мы можем задать простое правило классификации:

$$\forall\, j \quad x_j \in \Omega_j(k) \;\rightarrow\; x \in A_k, \;\; k \in \{1..K\} \qquad (5.1)$$

где $\Omega_j(k)$ - интервал значений признака x_j для класса A_k. Геометрическая интерпретация этого случая при n=2, K=3 показана на рис. 5.1.

На рис. 5.1 $\Omega_1(1)=\Omega_1(2)+\Omega_1(3)$, но класс A_1 разделяется попарно с классами A_2 и A_3 по признаку x_2. В то же время $\Omega_2(2)=\Omega_2(3)$, но эти классы разделяются по признаку x_1.

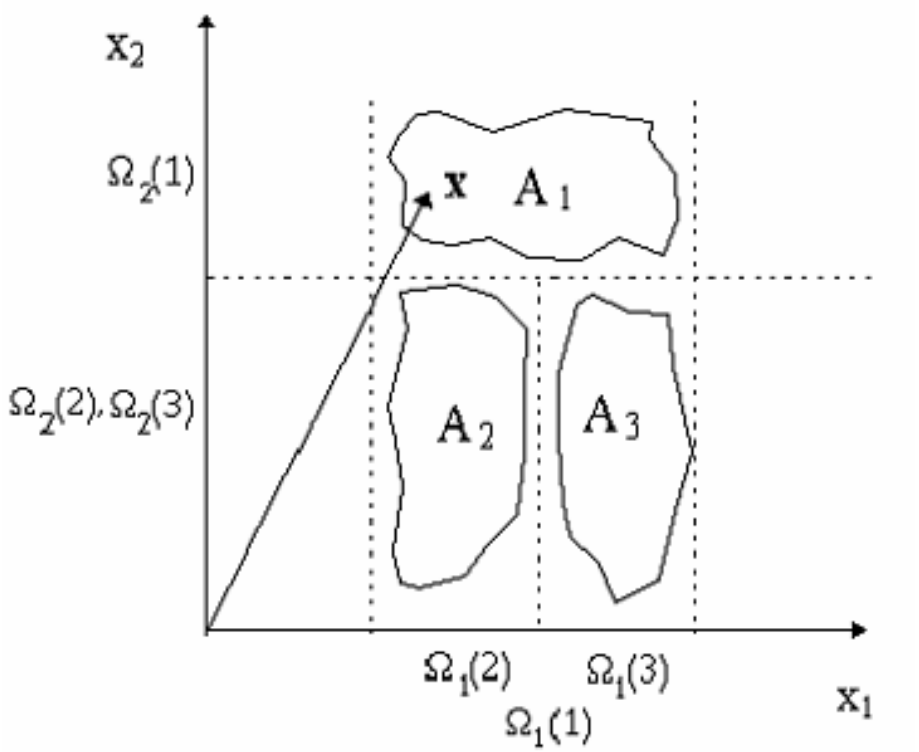

Рис.5.1. Простой случай разделения классов (гиперпараллелепипеды).

Метод классификации, основанный на правиле 5.1, называют **методом гиперпараллелепипедов**. Его разновидности имеются в большинстве пакетов обработки данных ДЗ. Но приведенный выше пример можно рассматривать только как идеальный частный случай. Реальные границы областей, соответствующих искомым классам в признаковом пространстве чаще всего имеют достаточно сложную форму; более того, эти области могут перекрываться для некоторых классов, особенно когда диапазоны значений параметров x_j невелики. Поэтому на практике возможности использования метода гиперпараллелепипедов ограничены.

В общем случае множество значений признаков для образов класса A_k - это область $\Omega(k)$ в пространстве **X**, ограниченная некоторой гиперповерхностью совершенно произвольной формы. Одному классу может соответствовать даже несколько таких областей. Перекрытие областей, соответствующих различным классам, приводит к ошибкам классификации. Поэтому во всех методах классификации образов, представленных n-мерными векторами данных, чаще всего априори вводятся некоторые предположения о характере группировки образов в пределах заданных классов. На основании этих предположений строятся гиперповерхности,

разделяющие множества Ω(k). Уравнения таких гиперповерхностей имеют вид d(**x**)=0 и называются **разделяющими функциями** или **дихотомиями.**

При отсутствии каких-либо специальных предположений о характере группировки образов границы областей Ω(k) можно аппроксимировать линейными функциями, то есть гиперплоскостями. Метод гиперпараллелепипедов - простейший случай такой аппроксимации, когда гиперплоскости параллельны координатным плоскостям. Более общий случай для n=2 представлен на рис. 5.2.

В случае n=2 разделяющие функции представляют собой уравнения прямых вида $d_i(\mathbf{x})=w_{i1}x_1+w_{i2}x_2+w_{i3}=0$. Функции $d_i(\mathbf{x})$ называют **линейными решающими функциями** или **дискриминантами.**

Разделяющая функция, заданная уравнением $d_i(\mathbf{x})=0$, делит пространство X на две области, для которых $d_i(\mathbf{x})>0$ и $d_i(\mathbf{x})<0$ соответственно. На рис.5.2 функция $d_1(x_1,x_2)$ положительна на множествах Ω(1) и Ω(2) и отрицательна на множествах Ω(3) и Ω(4), функция $d_2(x_1,x_2)$ - положительна на множестве Ω(1), функция $d_3(x_1,x_2)$ - на множествах Ω(1) и Ω(3). Отсюда можно записать систему неравенств, ограничивающую область решений для каждого класса. Например, на рис. 5.2, область решений в пользу класса A_1 задается неравенствами:

$$\begin{aligned} &d_1(x_1,x_2)>0,\\ &d_2(x_1,x_2)>0,\\ &d_3(x_1,x_2)>0. \end{aligned} \qquad (5.1)$$

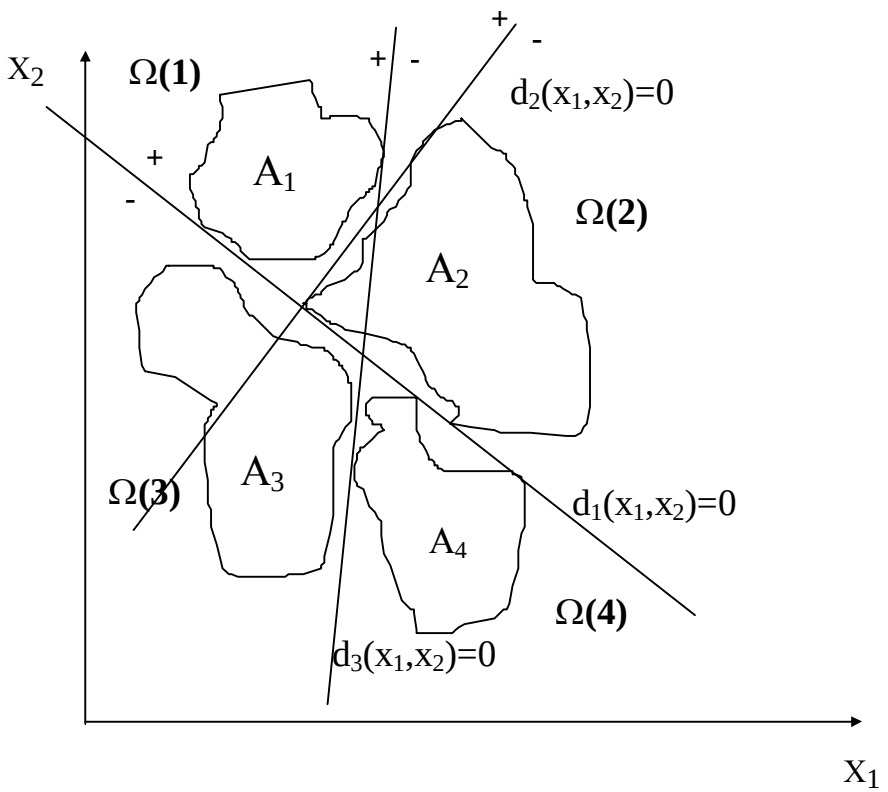

Рис.5.2. Разделяющие функции.

Если существует набор линейных решающих функций $\{d_i(\mathbf{x})\}$, с помощью которого можно задать область решения для каждого класса A_k, то классы $\{A_k\}$ называются линейно разделимыми. Задача построения такого набора заключается в отыскании коэффициентов w_{ij} $(j=1,...,n+1)$ каждой функции $d_i(\mathbf{x})$. Для этого обычно используется доступная выборка образов.

Заметим, что с помощью одних и тех же линейных решающих функций мы можем по-разному определить области решения для каждого класса. Так, при наборе ограничений (5.1) возникает так называемая "область отказов от распознавания" (клин между функциями d_2 и d_3 в верхней части рисунка), хотя этот набор более точно аппроксимирует границу класса A_1.

Можно обойтись без отказов, если определить область решения для A_1 только первыми двумя неравенствами.

Хотя приведенная выше запись решений удобна для геометрической интерпретации, аналитическое решение задачи удобнее искать, записав $d(\mathbf{x})$ отдельно для каждой пары соседних классов: $d_{12}(\mathbf{x})$, $d_{13}(\mathbf{x})$ и т.д. В этом случае уравнение разделяющих функций $d(\mathbf{x})_{ks}=0$ для пары классов k, s можно преобразовать к виду $r_k(\mathbf{x})-r_s(\mathbf{x})=0$ или $r_k(\mathbf{x})=r_s(\mathbf{x})$, а неравенства, ограничивающие область решения для k-го класса, представить в виде $r_k(\mathbf{x})>r_s(\mathbf{x})$. В такой постановке построение алгоритма распознавания при заданном наборе классов $\{A_k\}$ сводится к выбору К решающих функций $r_k(\mathbf{x})$, по которым на области $\Omega(k)$ выполняется условие

$$r_k(\mathbf{x})>r_s(\mathbf{x}) \quad \forall \quad s\neq k, \quad k,s=1,...,K. \qquad (5.2)$$

Тогда задача классификации конкретного вектора-образа **x** сводится к вычислению $r_k(\mathbf{x})$ для всех k=1,...,K и нахождению $\max\limits_k\{r_k(\mathbf{x})\}$.

В некоторых задачах, когда решающие функции имеют очень сложный вид, может быть использовано их обобщенное представление в виде

$$r_i(\mathbf{x}) = \sum_j w_{ij} f_j(\mathbf{x}), \qquad (5.3)$$

где $f_j(\mathbf{x})$ - некоторые известные функции. Далее, путем введения новых переменных $u_j= f_j(\mathbf{x})$, задача сводится к линейной.

В общем случае решающие функции далеко не всегда линейны. Например, при статистической классификации, которую мы будем в дальнейшем рассматривать, когда дисперсии признаков внутри классов различны, решающие функции представляют собой квадратичные формы (гиперквадрики). Тем не менее, представление решающих функций в линейной форме распространено достаточно широко в связи с простотой

автоматизации процесса обучения в таких системах и создания самообучающихся систем распознавания.

Процесс автоматического обучения в таких системах заключается в коррекции весовых коэффициентов w_{ij} функций (5.2) по результатам классификации предъявляемых системе векторов-образов. Для линейно разделимых классов существует целый ряд алгоритмов обучения, сходящихся за конечное число шагов, то есть позволяющих получить наилучшее приближение решающей функции на конечной выборке образов. Методология линейного дискриминантного анализа в настоящее время наиболее широко используется в нейронных сетях, где обучение является наиболее сложным и ответственным этапом процесса распознавания [31].

Непосредственное построение разделяющих или решающих функций без предположений о каких-либо закономерностях группировки образов внутри классов, которые можно описать математическими функциями, часто называют **непараметрическим** подходом. Применение непараметрического подхода целесообразно в тех случаях, когда нам ничего неизвестно о каких-либо закономерностях распределения образов внутри классов. При тематической обработке изображений земной поверхности такой метод бывает очень удобен для оценки положения определенных тематических категорий в признаковом пространстве. Простейший вариант этой процедуры реализуется в пакете ERDAS Imagine путем построения обучающих спектральных сигнатур непосредственно в признаковом пространстве (Selecting Feature Space Objects). Некоторые разновидности непараметрического подхода, применяющиеся при обработке изображений, рассмотрены в [11,12,18,19].

При большом количестве образов, классов и признаков-измерений непараметрическая классификация становится чрезвычайно сложной и чаще, в том числе и при поэлементной классификации изображений, используются методы, априори предполагающие существование некоторых закономерностей внутри классов. Такие методы мы и будем в дальнейшем рассматривать.

6. Группировка образов с использованием функций расстояния (кластерный анализ). Задачи кластерного анализа в тематической обработке данных ДЗ.

Кластером обычно называют группу образов $\{\mathbf{x}_i\}$, удовлетворяющих условию:

$$\|\mathbf{x}_i-\mathbf{x}_k\|<d, \qquad (6.1)$$

где ||.|| - мера сходства между образами,
d - заданное пороговое ограничение по этой мере.

Иногда кластеры называют **таксонами**, а кластерный анализ - **таксономией.**

Наиболее распространенной мерой сходства является расстояние между точками-образами в пространстве измерений (признаков) X. В большинстве случаев используется евклидова метрика

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{\sum_{i=1}^{n} (a_j^2 - b_j^2)} . \qquad (6.2)$$

В некоторых случаях могут использоваться и другие меры сходства, отличные от меры расстояния, например cos α, где α - угол между векторами **a** и **b** (мера корреляции). Этот подход лежит в основе так называемой **спектрально-угловой классификации**, обеспечивающей локальную инвариантность к условиям освещенности, которые часто необходимо учитывать при обработке изображений высокого разрешения [22]. В этом случае мера сходства не зависит от величины модуля вектора **x**, который характеризует яркость объекта в целом, а зависит только от соотношений значений яркости по каналам.

Кластеризация по корреляционной мере сходства в задаче распознавания имеет еще и другое, самостоятельное значение. Такая кластеризация применяется для сжатия сданных, в частности, для сокращения размерности признакового пространства. Этот класс задач называют **факторным анализом**. Он будет рассмотрен отдельно в соответствующем разделе.

Использование понятия кластера предполагает, что образы исследуемых объектов или явлений имеют естественную тенденцию к группировке вокруг некоторых характерных значений, которые называют **центрами кластеров**. Чем сильнее выражена эта тенденция, тем более успешно при решении задачи могут использоваться методы кластерного анализа.

При появлении многозональных космических сканеров кластерный анализ был одним из первых подходов, использованным при цифровой обработке многозональных сканерных изображений. Это объясняется тем, что при разрешении цифрового изображения несколько десятков или даже сотен метров на пиксел надежно разделяются только крупные элементы ландшафта подстилающей поверхности (водные объекты, лесные массивы, сельскохозяйственные угодья, открытые почвы, застройка, и т.п.). Эти объекты, как правило, хорошо различаются по спектральным характеристикам в том или ином спектральном диапазоне и достаточно однородны по яркости благодаря сглаженности сцены. При таких условиях

должна существовать тенденция к образованию групп в пространстве спектральных яркостных признаков **X**.

Недостатком такого способа классификации является отсутствие взаимосвязи системы координат яркостного пространства **X** с системой координат изображения. Выявленные кластеры далеко не всегда совпадают с теми объектами, которые интересуют обработчика. Поэтому методы кластерного анализа в пакетах обработки данных ДЗ называют **неконтролируемой** или **непомеченной** (английский термин **unsupervized**) классификацией.

Неконтролируемая классификация применяется обычно в следующих целях:

- для определения количества разделяющихся по спектральным характеристикам классов объектов на обследуемой территории;
- для оценки информативности имеющегося набора измерений при решении конкретной прикладной задачи;
- для выбора участков, которые могут быть использованы при формировании обучающих и контрольных выборок в процессе классификации с обучением.

При отсутствии данных наземных обследований или недостатке справочных (фондовых) материалов неконтролируемая классификация может, однако, оказаться единственным доступным способом тематической обработки. В этом случае стремятся выделить как можно большее количество кластеров и затем добиваются интерпретируемого результата, группируя кластеры уже с учетом расположения классифицированных точек на изображении.

Наиболее распространенные методы кластерного анализа можно условно разделить на две группы.

1. Методы выявления (выращивания) кластеров при заданном пороговом ограничении на расстояние между точками множества.
2. Методы формирования кластеров при заданном количестве групп.

В первом подходе количество кластеров, как правило, априори неизвестно. Исходными данными при такой постановке задачи являются пороговое ограничение расстояния d из (6.1) и правила объединения элементов множества. В результате количество и форма кластеров сильно зависят от выбранного метода анализа, величины порога и начальных условий.

По методам формирования кластеров в этом подходе выделяются односвязывающие методы (анализ элементов, ближайших к текущему), полносвязывающие методы (анализ наиболее удаленных элементов), и среднесвязывающие методы. Некоторые разновидности этих методов рассматриваются в монографиях [1,2,6], а также в [8,24].

Во втором подходе задается исходное количество центров кластеров, которые в процессе анализа перемещаются таким образом, чтобы заданные

требованиям к кластерам выполнялись наилучшим образом. Как правило, здесь имеется критерий качества кластеризации, который в процессе формирования кластеров максимизируется (или минимизируется).

К этой группе алгоритмов относятся алгоритмы К внутригрупповых средних, ISODATA (итеративный самоорганизующийся метод анализа данных) и дисперсионно-ковариационные критерии оптимизации. Алгоритмы данной группы представляют наибольший интерес, поэтому ниже мы рассмотрим все три типа таких алгоритмов.

В том и в другом подходе могут быть заданы дополнительные требования к кластерам: минимальное количество точек в группе, минимальное расстояние между группами и некоторые другие.

Существует большое количество более или менее эффективных методов кластеризации, основывающихся на теории графов. С некоторыми из них можно познакомиться в работах [9,24], однако в методологии тематической обработки обработки данных ДЗ такие алгоритмы не нашли применения.

6.1. Методы выявления (выращивания) кластеров.

Односвязывающий метод. Простейшим способом выявления кластеров является такой: выбираем произвольную точку (вектор) множества X, назначаем ее центром первого кластера ($\mathbf{x}_k$, k=1) и присоединяем к этому кластеру все точки, удовлетворяющие условию (6.1). Первая точка, для которой условие (6.1) не выполняется, назначается центром следующего кластера.

Далее для каждой точки вычисляется расстояние уже до двух центров, и она относится к тому кластеру, расстояние до которого меньше. Если расстояние до существующих кластеров больше заданного порога, образуется новый кластер и т.д. В конечном итоге мы получим некоторое разбиение на классы, вид которого, как уже говорилось, очень сильно зависит от порядка просмотра образов, особенно в тех случаях, когда тенденция к образованию групп прослеживается слабо. На рис. 6.1 показаны два варианта такого разбиения при различном порядке просмотра точек-образов. Легко показать, что разделяющие функции представляют собой геометрическое место точек, лежащих на перпендикулярах к отрезкам, соединяющим центры кластеров. Недостатком алгоритмов такого типа часто является образование так называемых "цепных" или "серпантинных" кластеров. Для образования компактных групп образов, близких к гиперсферам, удобнее использовать полносвязывающий метод, в котором выдение кластеров начинается с анализа наиболее удаленных точек.

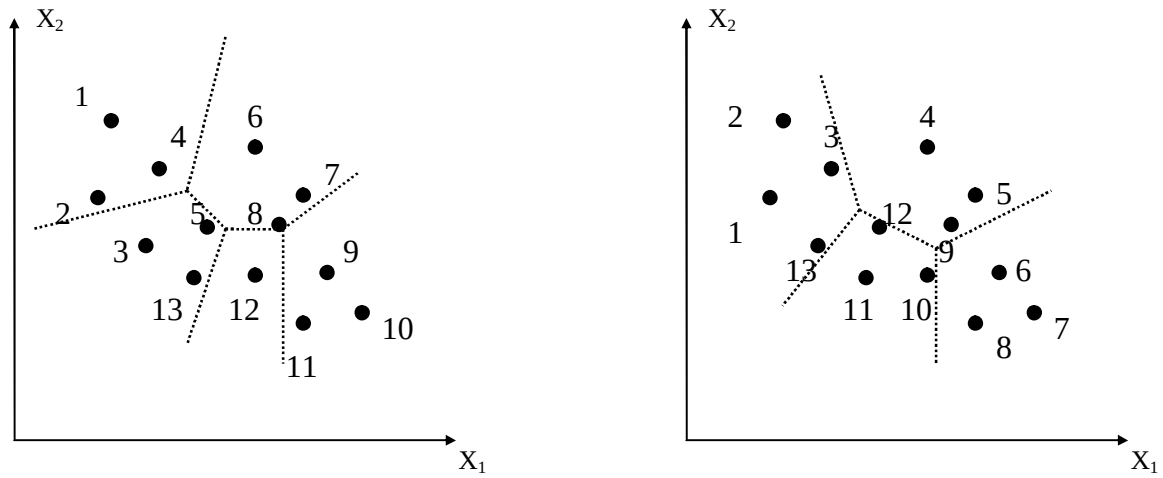

Рис. 6.1. Результаты кластеризации по порогу при различной последовательности анализа образов.

Полносвязывающий метод. Алгоритм максиминного расстояния. В качестве исходного образа выберем некоторую “крайнюю” точку, например, с минимальными координатами в пространстве X. Назовем ее центром $\mathbf{m_1}$ кластера K_1. В качестве второго центра K_2 выберем наиболее удаленную от нее точку по всему множеству образов. Определим пороговое значение d как

$$d=\|\mathbf{m_1}-\mathbf{m_2}\|/2. \qquad (6.2)$$

Шаг 1. Вычисляем расстояния до центров $\mathbf{m_1}$ и $\mathbf{m_2}$ $\|\mathbf{x}-\mathbf{m_1}\|$, $\|\mathbf{x}-\mathbf{m_2}\|$ для всех **x** нашего множества образов. Из каждой пары расстояний выбираем минимальное .

Шаг 2. Определяем максимальное значение

$$M=\max\{\min(\|\mathbf{x}-\mathbf{m_1}\|,\|\mathbf{x}-\mathbf{m_2}\|)\} \qquad (6.3)$$

по всему множеству образов. Пусть этому значению соответствует образ $\mathbf{x}_i$. Если M>d, назначаем $\mathbf{x}_i$ центром кластера K_3. В качестве новой пороговой меры d можно взять величину d=M/2 или половину среднего значения по всем минимальным расстояниям, рассчитанным на предыдущем шаге.

Шаг 3. Для всех **x** вычисляем минимальное из K расстояний до центров уже образованных кластеров: $\min\|\mathbf{x}-\mathbf{m_k}\|$, k=1,...,K.

Шаг 4. Вычисляем среднее минимальное расстояние образа **x** до центра кластера $\rho_{ср.}(\mathbf{x},\mathbf{m_k})=\| \mathbf{x}-\mathbf{m_k}\|/N$, где N - общее количество пар $(\mathbf{x},\mathbf{m_k})$. Назначаем новый порог $d=\rho_{ср.}(\mathbf{x},\mathbf{m_k})$.

Шаг 5. Ищем $\mathbf{x}_i$, соответствующий значению M из (6.3) по всему множеству образов. Если M<=d, процесс заканчивается. В противном случае назначаем $\mathbf{x}_i$ очередным центром кластера и переходим к шагу 3.

Процесс имеет смысл прекратить также в том случае, когда величина d становится меньше среднеквадратической погрешности измерений признаков, с которыми мы работаем.

Несмотря на то, что данный алгоритм позволяет получить компактные группы образов, он тоже является эвристическим, как и все алгоритмы выявления кластеров. Для получения более объективного результата необходимо использовать математически обоснованные показатели качества кластеризации. Эти критерии качества можно согласовать с целями конкретной задачи, для которой выполняется кластеризация.

6.2. Кластеризация при заданном количестве групп. Алгоритмы класса ISODATA.

Наиболее распространенным типом алгоритмов кластеризации, чаще всего имеющихся в пакетах тематической обработки данных ДЗ, являются алгоритмы, использующие как показатель качества минимум среднеквадратичной ошибки. Поэтому такие алгоритмы кластеризации называют иногда НСКО-алгоритмами, следуя терминологии, принятой в линейном дискриминантном анализе.

Заметим предварительно, что во многих случаях вместо евклидовой метрики используются просто квадраты расстояний, что, не влияя на результат, значительно упрощает расчеты.

Основой таких алгоритмов является кластеризация по методу **К внутригрупповых средних**. Исходными данными алгоритма К внутригрупповых средних является К центров кластеров $\mathbf{m}_1,...,\mathbf{m}_K$, заданных своими координатами в пространстве X. Эти К центров могут быть выбраны произвольно или рассчитаны каким-либо способом с учетом особенностей исходных данных. В процессе работы алгоритма происходит итеративная коррекция центров кластеров, минимизирующая сумму квадратов внутригрупповых расстояний.

Рассмотрим общую схему принятия решения на каждой итерации алгоритма.

Шаг 1. Все образы **x** распределяются по К кластерам по минимуму расстояния $\|\mathbf{x}-\mathbf{m_k}\|$.

Шаг 2. По результатам шага 1 рассчитываются новые центры кластеров $\mathbf{m_k}$. При этом предполагается, что сумма квадратов расстояний $\|\mathbf{x}-\mathbf{m_k}\|$ по всем $\mathbf{x}$ из k-го кластера должна быть минимальной. В этом случае центр $\mathbf{m_k}$ является выборочным средним для данного кластера:

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{k=1}^{N_k}\mathbf{x}_j, \quad (6.4)$$

где N_k-количество образов $\mathbf{x}$ в k-м кластере. Отсюда и название данного алгоритма.

Шаг 3. Если рассчитанные на шаге 2 новые центры совпадают с центрами, рассчитанными на предыдущей итерации (или отличаются от них не более, чем на заданную величину ε), процесс заканчивается. В противном случае переходим к шагу 1.

Несмотря на то, что алгоритм К средних давно и широко известен, доказательство сходимости было получено только в 1984 г. Селимом и Исмаилом [24]. Однако, как показывает практика (и это имеет теоретическое обоснование, также приведенное в [24]), величина I зависит от количества и положения исходных центров кластеров, а также и от свойств самого классифицируемого множества образов. Если в нем отсутствуют явно выраженные области сгущения точек, процесс кластеризации может затянуться надолго, и это необходимо учитывать при обработке больших массивов данных. Именно поэтому ограничение по числу итераций присутствует во всех компьютерных реализациях подобных алгоритмов.

Итеративный самоорганизующийся метод анализа данных (ISODATA - Itarative Self-Organizing Data Analysis Technique). Данный алгоритм, основанный на предыдущем методе, включает большое количество дополнительных параметров, что позволяет сделать процесс кластеризации более совершенным. На практике может использоваться какое-то подмножество параметров, но мы рассмотрим этот процесс в наиболее общем виде [1].

Процедура кластеризации включает три основных блока: 1) блок аппроксимации центров; 2) блок расщепления центров; 3) блок слияния центров. В отдельный блок можно также вынести процедуру, осуществляющую передачу управления при завершении одного из указанных процессов.

Для работы такого алгоритма необходимо задать следующие параметры.

K - исходное количество кластеров,

($\mathbf{m}_1$,...,$\mathbf{m}_K$) - набор исходных центров,

N - желаемое количество количество кластеров,

θ_s- ограничение снизу на количество образов в кластере,

θ_σ - ограничение снизу на внутригрупповое среднеквадратичное отклонение в блоке расщепления кластеров,

θ_c- параметр компактности - ограничение сверху на расстояние между центрами кластеров для блока слияния,

L - максимальное количество пар центров, которые можно объединить в блоке слияния,

I_{max} - допустимое количество итераций.

1. Аппроксимация центров.

Шаг 1. Распределяем все образы **x** по K кластерам по минимуму расстояния до центров:

$$\|\mathbf{x}-\mathbf{m}_p\|=\min\|\mathbf{x}-\mathbf{m}_k\|,\ k=1,...,K \ \rightarrow \ \mathbf{x}\in\Omega p.$$

Шаг 2. Проверяем размеры образованных кластеров и исключаем слишком мелкие по условию:

$N_k<\theta_s \ \Rightarrow$ удаляем $\mathbf{m}_k$ и уменьшаем количество центров: K=K-1.

Здесь N_k - число образов в k-м кластере.

Если на этом шаге количество кластеров сократилось, возвращаемся к шагу 1, чтобы распределить образы по оставшимся кластерам.

Шаг 3. Рассчитываем новые центры кластеров как выборочные средние по формуле (6.4).

Шаг 4. Вычисляем для образов каждого кластера среднее расстояние до центра

$$D_{cp.}(k)=\frac{1}{N_k}\Sigma\|\mathbf{x}_i-\mathbf{m}_k\|,\ i=1,..., N_k,\ k=1,...,K. \qquad (6.5)$$

Шаг 5. Вычисляем обобщенное среднее расстояние по всем кластерам:

$$D=\frac{1}{K}\sum_{k=1}^{K}N_k D_{cp.}(k). \qquad (6.6)$$

Ветвление процесса.

Шаг 6.

1) если количество итераций $I=I_{max}$, то задаем $\theta_c=0$ и переходим к **шагу 12** (завершение процесса без слияния кластеров) ;

обнуление θ_c обеспечивает автоматический выход из блока слияния, так как в этом случае мы не сможем найти ни одной пары кластеров для объединения;
2) если $K \le N/2$, то переходим к **шагу 7** (расщепление кластеров);
3) если **I** - четно и $K \ge 2N$, то переходим к **шагу 12** (слияние кластеров);
4) если не выполнено ни одно из условий 1-3, то подставляем новые значения центров $\mathbf{m}_k$ и переходим к **шагу 1** (продолжаем аппроксимацию центров).

2. Расщепление центров.

В этом блоке выполняется расщепление наиболее «растянутых» в каком-то направлении кластеров, для чего предварительно оценивается, насколько они растянуты и в каком именно направлении.

Шаг 7. Для каждого из K кластеров рассчитывается вектор среднеквадратичного отклонения $\sigma_{\mathbf{k}}$. Компоненты этого вектора по координатам n-мерного пространства X вычисляются по формуле

$$\sigma_{jk}=\sqrt{\frac{1}{N_k}\sum_{x\in\Omega_k}(x_{ik}-m_{ik})^2},\ k=1,...,K,\quad j=1,...,n \qquad (6.7).$$

Шаг 8. Для каждого кластера k=1,...,K определяется максимальная компонента по всем координатам пространства X (то есть направление максимального растяжения)

$$\sigma_{max}(k)=\max \sigma_{jk},\ j=1,...,n.$$

Шаг 9. Если для всех выбранных $\sigma_{max}(k)$, k=1,...,K выполняется условие $\sigma_{max}(k) > \theta_\sigma$, то переходим к следующему шагу, иначе переходим к **шагу 12**.

Шаги 10 и 11 выполняются в цикле по всем кластерам k=1,...,K.

Шаг 10. Проверяется выполнение условий
а) $D_{cp.}(k) > D$ и $N_k > 2(\theta_s+1)$;
б) $K \le N/2$.

Если хотя бы одно из этих условий для k-го кластера выполняется, переходим к шагу 11, иначе k=k+1 (переходим к проверке следующего кластера).

Шаг 11. Расщепляем k-й кластер, для которого выполнены условия шага 10.

Расщепление производится по координате j, соответствующей максимальной компоненте $\sigma_{max}(k)$. Для этого выбирается некоторая величина

$0<\gamma\leq 1$ и из центра $\mathbf{m_k}$ формируются два новых центра, отличающиеся по координате j:

$$m_{jk}^{+}=m_{jk}+\gamma\sigma_{max}(k),\ m_{jk}^{-}=m_{jk}-\gamma\sigma_{max}(k).$$

Величина γ выбирается так, чтобы новые кластеры были различимы, но не повлияли существенно на положение других кластеров.

3. Слияние центров.

Шаг 12. Вычисляются расстояния D_{ks} между всеми парами кластеров:

$$D_{kl}=\|\mathbf{m}_k-\mathbf{m}_l\|\ ,\ k=1,...,K,\ l=1,...,K,\ k\neq s.$$

Шаг 13. Для всех пар проверяется выполнение условия $D_{kl} < \theta_c$. Те пары, для которых оно выполняется, упорядочиваются по возрастанию D_{ks} и выбираются первые L пар - кандидатов на слияние. Если таких пар нет, переходим к **шагу 15**.

Шаг 14. Слияние происходит попарно. Новый центр для каждой из L пар вычисляется по формуле:

$$\mathbf{m}_p^*=\frac{1}{N_k+N_l}(N_k\mathbf{m}_k+N_l\mathbf{m}_l).$$

Центры $\mathbf{m}_k$ и $\mathbf{m}_l$ ликвидируются и количество кластеров уменьшается: K=K-1.

Шаг 15. Если I=Imax, то процесс заканчивается. В противном случае переходим к шагу 1.

Описанный алгоритм в его классическом виде целесообразно применять только на небольших объемах данных. Для обработки очень больших объемов данных, в частности, для многозональных изображений обычно используются различные усеченные версии. Иногда вместо алгоритма К средних используется **алгоритм Форджи** [19], в частности, и в пакете ERDAS Imagine. Алгоритм Форджи отличается от алгоритма К средних тем, что сходимость процесса кластеризации оценивается не по расстоянию между центрами кластеров i-1–й и i-й итераций, а по порогу согласованности (convergence threshold), определяющем процент точек кластера, который не переместился из одного кластера в другой между двумя итерациями. В пакете ERDAS Imagine задаются только необходимое количество классов и порог согласованности. В пакете IDRISI задаются исходное и желаемое количество классов. Допустимое количество итераций задается во всех реализациях.

Безусловно, некоторые из параметров могут быть заданы внутри самого алгоритма или рассчитываться на основе статистических оценок по анализируемому множеству данных [24,30]. Однако нельзя забывать, что алгоритмы кластеризации относятся к процедурам эвристического типа и как само число исходных параметров, так и способ их задания неизбежно отражаются на результате. Достаточно устойчивый по отношению к исходным параметрам результат мы можем получить только в тех случаях, когда на множестве образов четко прослеживаются области сгущения. Если мы построим многомерную гистограмму для этого множества, то она будет характеризоваться некоторым количеством ярко выраженных пиков (мод), которые и будут соответствовать наиболее устойчивым кластерам. Если наша задача заключается в поиске именно таких кластеров, то мы получим хороший результат при любом способе задания исходных центров. Но в задачах классификации изображений особенно высокого разрешения, тенденция к группировке образов вокруг каких-то центров может вообще не наблюдаться, и в этих случаях, задав исходные центры, мы, фактически, выполняем обыкновенную классификацию по минимуму расстояния, основанную на принципе сравнения с эталоном. В таком случае способ задания центров кластеров становится для этой процедуры решающим [16,30].

6.3. Дисперсионно-ковариационные критерии качества кластеризации.

В рассмотренных ранее алгоритмах критерием качества кластеризации является минимум среднего внутригруппового растояния элементов кластера до его центра. Это расстояние мы рассчитывали по формуле (6.5). Однако, как уже говорилось, в алгоритмах кластеризации вместо среднего евклидова расстояния чаще используется сумма квадратов расстояний [6,24], чтобы упростить расчеты и устранить зависимость функционала качества кластеризации от нормировочных коэффициентов $1/N_k$. То есть вместо (8.3) обычно используется выражение

$$D^2(\mathbf{x}_i,\mathbf{m}_k)=(\mathbf{x}_i-\mathbf{m}_k)^T(\mathbf{x}_i-\mathbf{m}_k),\ i=1,...,N,\ k=1,...,K. \quad (6.8)$$

Здесь $N=\Sigma N_k$ - общее количество образов в пространстве X, K - количество кластеров, - центр k - го кластера, рассчитанный по формуле (6.4). $\mathbf{d}=\mathbf{x}_i-\mathbf{m}_k$ - это вектор, соединяющий центр $\mathbf{m}_k$ с точкой $\mathbf{x}_i$. Скалярному произведению $\mathbf{d}^T\mathbf{d}$ соответствует матрица $\mathbf{M}=\mathbf{d}\mathbf{d}^T$, которую называют матрицей рассеяния или матрицей разброса. Для k-го кластера матрица внутригруппового рассеяния имеет вид:

$$\mathbf{M}_w(k)= \sum_i (\mathbf{x}_i - \mathbf{m}_k) (\mathbf{x}_i - \mathbf{m}_k)^T,\ i=1, ..., N_k , k=1,...,K. \quad (6.9)$$

Величина $D^2(\mathbf{x}_i,\mathbf{m}_k)$ представляет собой сумму диагональных элементов матрицы $\mathbf{M}_w(k)$ и обозначается tr $\mathbf{M}_w(k)$ (след матрицы).

Используя введенные понятия, можно получить некоторые полезные соотношения, характеризующие структуру кластеров в пространстве X. Эти соотношения используются для формирования различных функционалов качества кластеризации, которые называют дисперсионно-ковариационными критериями [2,7].

Определим обобщенную матрицу внутригруппового рассеяния следующим образом:

$$\mathbf{M}_w=\frac{1}{N}\sum_k\sum_i (\mathbf{x}_i - \mathbf{m}_k) (\mathbf{x}_i - \mathbf{m}_k)^T,\ i=1, ..., N_k , N=\sum_k N_k , k=1,...,K. \quad (6.10)$$

Величина tr $\mathbf{M}_w/K$ представляет собой усредненную сумму квадратов расстояний по всем К кластерам, то есть $\mathbf{M}_w$ характеризует средний разброс элементов вокруг центров в выделенных группах.

Аналогично введем матрицу разброса между кластерами.

$$\mathbf{M}_b=\sum_k \frac{N_k}{N} (\mathbf{m}_k-\mathbf{m}) (\mathbf{m}_k-\mathbf{m})^T,\ k=1,...,K. \quad (6.11)$$

Здесь **m** - обобщенный центр, соответствующий выборочному среднему по всему
множеству образов:

$$\mathbf{m}=\frac{1}{N}\sum_i \mathbf{x}_i =\frac{1}{K}\frac{1}{N_k}\sum_j \mathbf{x}_j,\ i=1,..., N_k , N=\sum_k N_k , j=1,..., N_k . \quad (6.12)$$

Наконец, определим матрицу рассеяния по всему множеству образов:

$$\mathbf{M}_0=\frac{1}{N}\sum_i (\mathbf{x}_i - \mathbf{m}) (\mathbf{x}_i - \mathbf{m})^T,\ i=1, ..., N. \quad (6.13)$$

След матрицы $\mathbf{M}_0$ называют статистическим рассеянием множества образов.

Для матриц (6.11)-(6.13) имеет место следующее важное соотношение:

$$\mathbf{M}_0= \mathbf{M}_b + \mathbf{M}_w. \quad (6.14)$$

Доказательство данного соотношения достаточно простое и приводится, например, в [24]. Из этого соотношения следует, что на

множестве образов с известной матрицей рассеяния при минимизации внутригруппового рассеяния максимизируется межгрупповое рассеяние. То есть, сокращая среднее внутригрупповое расстояние, мы одновременно увеличиваем среднее расстояние между кластерами.

Из-за частичного отбрасывания (а на практике иногда и отсутствия) нормировочных коэффициентов введенные матрицы рассеяния часто называют матрицами сумм квадратов.

В алгоритмах кластеризации встречаются и другие функционалы качества, использующие матрицы рассеяния. Например, инвариантными к невырожденным линейным преобразованиям пространства X являются функционалы:

$$\beta=\mathrm{tr}\ \mathbf{M}_w{}^{-1}\mathbf{M}_b;$$
$$\beta=\mathrm{tr}\ \mathbf{M}_0{}^{-1}\mathbf{M}_w;$$
$$\beta=|\mathbf{M}_w|/|\mathbf{M}_0|.$$

Могут использоваться также функционалы $\beta=\mathrm{tr}\ \mathbf{M}_b/\mathrm{tr}\ \mathbf{M}_w$, $\beta=\ln\{|\mathbf{M}_w+\mathbf{M}_b|/|\mathbf{M}_w|\}$ и некоторые другие. Примеры практического использования некоторых функционалов качества при обработке изображений можно найти в [21,29].

В системах тематической обработки аэрокосмических многозональных изображений особенность кластеризации состоит в том, что выделяемые в пространстве X кластеры должны быть интерпретируемы, то есть образовывать на изображении связные области, соответствующие определенным тематическим категориям. Привлекательным с этой точки зрения качеством алгоритмов, минимизирующих внутригрупповой разброс, является более высокая пространственная однородность полученных классов объектов и более четкие границы между ними, чем при некоторых других способах классификации. Это упрощает интерпретацию и формирование тематической карты. Однако с точки зрения информативности результата они имеют ряд недостатков.

Заметим, что одинаковые средние значения введенных характеристик разброса могут получаться при различных положениях центров кластеров, то есть достигнутый минимум нашего функционала не гарантирует единственность решения (то есть не является глобальным минимумом, [24]). Следовательно, задача может быть полностью решена только путем перебора различных вариантов начальных условий. Но при произвольном задании центров в пространстве яркостных признаков соотношение (6.14) часто приводит к нежелательным последствиям. Причиной этого являются различия яркостных диапазонов каналов видимой и ИК зон спектра. Более зашумленные каналы видимой зоны имеют в 2-3 раза меньший ненулевой диапазон яркостей, чем ближний ИК-диапазон. В результате кластеры

вытягиваются вдоль наиболее информативной компонеты, что приводит к потере информации из других каналов.

В пакете ERDAS Imagine центры в автоматическом режиме задаются по линейной схеме, то есть располагаются либо по диагонали гиперкуба, образованного осями координат пространства X, либо по наиболее информативному направлению (максимальному диаметру множества X). Обычно это приводит к тому, что наибольшее дробление на кластеры происходит по почвенно-растительным классам, а такие важные для многих приложений тематические категории, как вода, заболоченные участки и техногенные загрязнения, имеющие в ближнем ИК-диапазоне низкую отражательную способность, разделяются плохо даже при большом количестве исходных центров или вообще не разделяются. Причем расположение центров по диагонали гиперкуба часто позволяет получить в подобных ситуациях лучший результат, чем при выборе центров по главной компоненте.

В пакете IDRISI выбор центров производится путем квантования результата RGB-синтеза, что в некоторых случаях приводит к более удачному результату. Тем не менее, количество кластеров там, как правило, сильно ограничено, поэтому часть информации неизбежно пропадает.

Поэтому при использовании НСКО-алгоритмов кластеризации, в том числе и ISODATA, необходимо предварительно выбрать подходящую методику задания исходных центров, каким-то образом связанную с реальными объектами изображения. Одним из способов может быть предварительная сегментация изображения на однородные области и выбор исходных центров по группе сегментов, наиболее однородных по яркостному признаку во всех каналах [15,16]. Другой способ, также с использованием пространственной статистической информации, рассмотрен в [30]. При таком выборе исходных центров результат кластеризации наиболее полно для заданного количества кластеров сохраняет границы объектов на изображении и значительно облегчает тематическую интерпретацию результата.

Контрольные вопросы к разделам 5-6.

1. Дайте определение разделяющей и решающей функции. Как определяется принадлежность образа к классу 1) через разделяющие функции; 2) через решающие функции.
2. Что такое области отказов от распознавания? Как они возникают?
3. Приведите примеры непараметрического и параметрического подходов к классификации образов, представленных набором параметров.
4. В каких случаях кластеризация только по пороговому ограничению может дать удовлетворительный результат?

5. Назовите основные этапы кластеризации по методу ISODATA. Какие параметры используются на каждом из этапов? Чем отличается алгоритм Форджи от алгоритма ISODATA?

6. Какой показатель (функционал) максимизируется алгоримами класса ISODATA? Дайте обоснование этого факта через матрицы сумм квадратов.

7. Перечислите основные достоинства и недостатки методов кластерного анализа для задач тематической обработки аэрокосмической информации.

7. Выбор признаков. Корреляционный анализ данных.

Системы дистанционных измерений, особенно космические, обычно рассчитаны на широкий спектр прикладных задач. И если для выделения конкретных классов объектов мы будем использовать все доступные наборы измерений (например, все каналы Тематического Картографа), мы можем получить задачу слишком большой размерности. Это затруднит если не сами расчеты, то, по меньшей мере, получение представительных выборок для описания классов в системе координат признаков. Поэтому после выбора перечня классов можно подумать над тем, как описать эти классы с использованием меньшего количества измерений.

В целом в задачах распознавания образов преобразование признакового пространства **X** обычно имеет две цели:

1) сократить общее количество признаков n, сохранив при этом максимальное количество полезной информации;
2) сделать эти признаки статистически независимыми, то есть образующими ортогональный базис в пространстве **X**.

Статистическая независимость признаков значительно упрощает расчеты при описании классов в пространстве **X** и непосредственно при классификации.

Предположим, что значения координат образов в признаковом пространстве **X** для каждого из наших классов распределены по нормальному (гауссову) закону. Эта гипотеза часто используется, в частности, при классификации аэрокосмических изображений, так как на измерения коэффициентов отражения подстилающей поверхности воздействует множество случайных факторов различной природы.

В общем случае плотность многомерного нормального распределения имеет вид

$$p(\mathbf{x})=\{1/[(2\pi)^{n/2}|\mathbf{C}|^{1/2}]\}\exp[-1/2(\mathbf{x}-\mathbf{m})^{T}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})],$$

где **m** – среднее по множеству **X**, а **C** – ковариационная матрица размерности n x n, элементы которой определяются как

$$\sigma_{ij}=M[(x_{ij}-m_i)(x_{ij}-m_j)],\ i=1,...,n,\ j=1,...,n. \quad (7.1)$$

Ковариационная матрица – это, собственно, та же матрица разброса образов в классе относительно центра распределения **m** (6.13). Элементы этой матрицы отражают взаимосвязь между координатами векторов из множества **X**. Если для системы координат измерений - базиса пространства **X** - выполняется условие ортогональности, то есть скалярные произведения $(\mathbf{x}_i,\mathbf{x}_j)=0$, $i=1,...,n$, $j=1,...,n$, это означает статистическую независимость параметров x_j, $j=1,...,n$. В этом случае **C** будет диагональной матрицей с элементами $\sigma_{ii}=(\sigma_i^2)$, которые представляют собой дисперсии множества по параметрам x_i. Для описания распределения вместо матрицы **C** тогда достаточно иметь n значений координат вектора σ.

В интерактивных системах обработки многозональных изображений, в частности, для выбора эталонных участков при обучении классификаторов, ортогональное преобразование признакового пространства иногда используется с целью визуализации тех классов объектов, которые не различаются в исходной системе координат пространства признаков, хотя и могут быть разделены автоматическими методами классификации (рис.7.1).

Диапазоны значений признаков $\mathbf{x}_1$ и $\mathbf{x}_2$ для классов A и B, изображенных на рис .7.1, перекрываются, поэтому при визуальном анализе изображения, когда определенные цвета присваиваются определенным интервалам, нам не удастся различить эти классы в исходной системе координат, хотя ясно, что для них можно построить разделяющую функцию (в данном примере совпадающую с осью координат $\mathbf{f}_1$). В этом случае преобразование признакового пространства к системе координат $\mathbf{f}_1,\mathbf{f}_2$ позволит нам «увидеть» объекты, принадлежащие к этим классам, так как они различаются по признаку $\mathbf{f}_2$.

Анализ многомерных статистических данных с использованием преобразования системы координат исходных параметров к новому ортогональному базису обычно называют **анализом главных компонент**.

Мы уже говорили о том, что взаимосвязь между координатами измерений отражают элементы матрицы внутригруппового рассеяния или, иначе, ковариационной матрицы. При анализе главных компонент иногда используют непосредственно ковариационную матрицу (7.1), но в наиболее общем случае - корреляционную матрицу. Рассмотрим, что представляет собой эта матрица, более подробно.

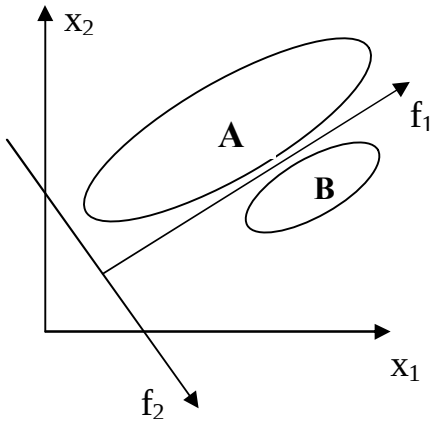

Рис.7.1. Разделение классов путем преобразования к главным компонентам.

Корреляционная мера и корреляционная матрица. Корреляционная мера сходства двух векторов **a** и **b** определяется как косинус угла между этими векторами:

$$\rho(\mathbf{a},\mathbf{b})=(\mathbf{a},\mathbf{b})/(\|\mathbf{a}\|\cdot\|\mathbf{b}\|), \qquad (7.2)$$

где **(a,b)** - скалярное произведение векторов, $\|\cdot\|$ - норма (длина) вектора.

Ясно, что величина ρ будет принимать значения на отрезке [-1,1], при этом она будет положительна при одинаковых знаках соответствующих координат векторов **a** и **b** и отрицательна при противоположных. При конечном числе образов N эта величина есть ни что иное, как выборочный коэффициент корреляции по наборам из i=1,...,N реализаций случайных величин ξ_j и ξ_k.

При значении меры корреляции ρ_{jk}, близкой по модулю к 1, в некоторых случаях можно практически без потери информации использовать при распознавании только один из признаков x_i или x_j. Тем не менее, перед этим все-таки целесообразно проверить, какие классы разделяются по этому признаку и не является ли наличие данного измерения принципиально важным для нашей конкретной задачи.

Матрица $\mathbf{R}=\{\rho_{jk}\}$, j=1,...,n, k=1,...,n размерности n x n называется **корреляционной матрицей**. Из определения коэффициента корреляции ясно, что корреляционная матрица R - симметрическая положительно полуопределенная, с диагональными элементами $\rho_{jj}=1$, j=1,...,n.

Наиболее удобным формализованным способом расчета корреляции между признаками-измерениями по выборке образов является использование так называемой стандартизованной матрицы данных [5]. Эта схема удобна и при расчете корреляции между атрибутивными описаниями объектов в ГИС, например, при решении некоторых задач картографической генерализации.

Стандартизованная матрица данных. Пусть у нас имеется N образов, представляющих собой векторы в n-мерном пространстве измеряемых признаков - реализаций n-мерной случайной величины ξ. Представим полученные данные в виде матрицы размерности n x N.

$$X^0 = \begin{pmatrix} x^0_{11} & \dots & x^0_{1j} & \dots & x^0_{1n} \\ \vdots & & \vdots & & \vdots \\ x^0_{i1} & \dots & x^0_{ij} & \dots & x^0_{in} \\ \vdots & & \vdots & & \vdots \\ x^0_{N1} & \dots & x^0_{Nj} & \dots & x^0_{Nn} \end{pmatrix}, \quad (7.3)$$

i=1,...,N, j=1,...,n.

Заметим, что каждый столбец матрицы - это вектор в пространстве размерности N, где N –число наблюдений (образов). В таком N-мерном пространстве задачу выделения наиболее информативных признаков (снижения размерности) можно рассматривать как задачу кластеризации по корреляционной мере сходства, о чем уже упоминалось в разделе 6.

Для решения этой задачи приведем матрицу данных $\mathbf{X}^0$ к стандартизованному виду.

1. Рассчитаем выборочные средние по каждой компоненте (столбцу) j=1,...,n:

$$m_j = \frac{1}{N}\sum_{i=1}^{N} x^0_{ij} \quad . \qquad (7.4)$$

2. Вычислим выборочную дисперсию по каждой компоненте:

$$(\sigma^0_j)^2 = \frac{1}{N}\sum_{i=1}^{N}(x^0_{ij} - m^0_j)^2. \quad (7.5)$$

Дисперсия, рассчитанная по (7.3), является смещенной оценкой с точки зрения математической статистики, но здесь она рассматривается скорее как среднее внутригрупповое расстояние, подобно тому, как это делалось в алгоритмах кластеризации.

Элементы стандартизованной матрицы данных $\mathbf{X}=\{x_{ij}\}$ вычисляются по формуле:

$$x_{ij} = \frac{x^0_{ij} - m^0_j}{\sigma^0_j}. \quad (7.6)$$

Стандартизованную матрицу данных **X** иначе называют **нормированной матрицей**. В результате проведенных операций мы перемещаем начало координат пространства признаков в точку с координатами $m_1,...,m_n$ и нормируем шкалу по каждой координате на значение σ. Полученная таким образом стандартизованная матрица данных обладает следующими свойствами.

1. mj=0 для всех j=1,...,n.
2. $\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1.$
3. $\|x_i\| = \|x_j\| = \sqrt{N}.$

Мера корреляции (7.2) между двумя измеряемыми параметрами j и k, представленными вектор-столбцами матрицы **X**, тогда принимает вид:

$$\rho_{jk}=(1/N)(\mathbf{x}_j,\mathbf{x}_k). \quad (7.7)$$

Заметим, что $\rho_{jk}=\sigma_{ij}/(\sigma_i\sigma_j)$, где σ_{ij} - выборочная ковариация между случайными величинами ξ_j и ξ_k. Именно поэтому при выполнении ортогонального преобразования к главным компонентам иногда пользуются не корреляционной, а ковариационной матрицей (7.1). Тем не менее, не следует путать эти два понятия.

7.1. Элементы факторного анализа. Метод главных компонент.

Задачи корреляционного и факторного анализа возникли значительно раньше методологии распознавании образов. Чаще всего они встречаются в естественнонаучных исследованиях при выявлении факторов (параметров, характеристик), наиболее существенно влияющих на тот или иной процесс, или их комбинаций. Отсюда и название факторный анализ. В геоинформационных технологиях факторный анализ может применяться при картографической генерализации и при создании так называемых синтетических карт, когда производится классификация таксономических единиц карты на несколько градаций по целому комплексу показателей.

Именно потому, что факторный анализ имеет своей основной целью снижение размерности множества данных, мы рассмотрим эту задачу в наиболее общем виде, где ортогональность новых параметров является всего лишь дополнительным условием.

Общая постановка задачи. Пусть у нас есть n параметров (факторов), заданных векторами $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n$. Требуется выбрать m<n наиболее значимых (важных) факторов $\mathbf{f}_1,\mathbf{f}_2,...,\mathbf{f}_m$.

Заметим, что здесь исходными параметрами являются вектор-столбцы стандартизованной (нормированной) матрицы данных, элементы которой представлены в виде (7.6).

В задачах факторного анализа решение ищется из системы линейных уравнений вида

$$\mathbf{x}_j=\sum_{k=1}^{m} \alpha_{kj}\mathbf{f}_k+\xi_j, \quad j=1,...,n. \qquad (7.8)$$

Коэффициенты α_{kj} называются **факторными нагрузками**, а свободный член ξ_j - **характерным фактором**.

Уравнение (7.8) отличается от обычной линейной регрессии тем, что заранее неизвестны ни сами факторы $\mathbf{f}_k$, ни факторные нагрузки α_{kj}. Фактор ξ_j играет роль шумовой компоненты, поэтому $M(\xi_j)=0$, но дисперсия также заранее неизвестна.

Различные модели факторного анализа отличаются критериями близости между вычисленными и измеряемыми параметрами, а также рядом дополнительных условий [5]. Мы будем рассматривать **модель главных компонент**, в которой к условию (7.8) добавляется условие ортогональности:

$$(\mathbf{f}_k,\mathbf{f}_l)=\begin{cases} N, & k = l \\ 0, & k \neq l \end{cases}, \quad k=1,...,n; \quad l=1,...,n. \qquad (7.9)$$

Векторы

$$\hat{\mathbf{x}}_j = \sum_{k=1}^{m} \alpha_{kj}\mathbf{f}_k = \mathbf{x}_j-\xi_j \qquad (7.10)$$

называются **вычисленными параметрами**.

В модели главных компонент критерием близости является минимальная суммарная дисперсия характерных факторов (то есть, как и в НСКО-алгоритмах кластеризации, минимизируется сумма квадратов ошибок).

$$\sum_{j=1}^{n} \sigma^2(\xi_j)=\frac{1}{N}\sum_{j=1}^{n} (\xi_j, \xi_j) \to \min. \qquad (7.11)$$

Рассматривая это выражение как функцию факторных нагрузок и дифференцируя его по α_{kj}, можно получить следующую систему уравнений:

$$\frac{\partial F(\alpha)}{\partial \alpha_{lj}} = -\frac{2}{N}\sum_{j=1}^{n} (\mathbf{x_j},\mathbf{f}_l)+ \frac{2}{N}\sum_{j=1}^{n} \sum_{k=1}^{m} \alpha_{kj}(\mathbf{f_k},\mathbf{f}_l). \qquad (7.12)$$

Для того, чтобы определить, при каких условиях будет выполнено требование (7.11), приравняем полученные производные нулю и после небольших преобразований придем к выражению:

$$\sum_{j=1}^{n} (\mathbf{x_j} - \sum_{k=1}^{m} \alpha_{kj}\mathbf{f_k},\mathbf{f_l})=0. \qquad (7.13)$$

С учетом (7.10) из этого выражения следует, что

1) характерные факторы ортогональны общим факторам: $(\xi_j,\mathbf{f}_l)=0$;

2) α_{lj} – это проекции параметров $\mathbf{x_j}$ на общие факторы $\mathbf{f}_l$: $\alpha_{lj} = \frac{1}{N}(\mathbf{x_j},\mathbf{f}_l) = \rho(\mathbf{x_j},\mathbf{f}_l)$.

Из ортогональности характерных факторов вычисленным параметрам следует, что $\sigma^2(\mathbf{x}_j) = \sigma^2(\hat{x}_j)+\sigma^2(\xi_j)$. Так как для нормированных исходных параметров $\sigma^2(\mathbf{x}_j)=1$, то $\sum_{j=1}^{n} \sigma^2(\mathbf{x}_j)=n=\text{const}$, $\sum_{j=1}^{n} \sigma^2(\hat{x}_j)=n-\sum_{j=1}^{n} \sigma^2(\xi_j)$, откуда следует, что минимизация функционала (7.11) эквивалентна нахождению максимума функционала $\sum_{j=1}^{n} \sigma^2(\hat{x}_j)=\sum_{j=1}^{n}\sum_{k=1}^{m} \alpha_{kj}$. А поскольку $\alpha_{lj} = \rho(\mathbf{x_j},\mathbf{f}_l)$, то в конечном итоге мы будем максимизировать корреляцию между исходными нормированными параметрами и искомыми общими факторами

$$\sum_{j=1}^{n}\sum_{k=1}^{m} \rho(\mathbf{x_j},\mathbf{f}_k) = \frac{1}{N}\sum_{j=1}^{n}\sum_{k=1}^{m} (\mathbf{x_j},\mathbf{f}_k)^2 \text{ -> max.} \qquad (7.14)$$

Схема решения модели главных компонент.

Отбросив в выражении (7.11) множитель $\frac{1}{N}$ и изменив порядок суммирования, переходим к функционалу:

$$\sum_{k=1}^{m}\sum_{j=1}^{n}(\mathbf{x}_j,\mathbf{f}_k)^2 \to \max . \qquad (7.15).$$

После этого можно искать максимум данного аддитивного функционала, последовательно находя максимум каждого из слагаемых. То есть решение задачи будет осуществляться по следующей схеме.

Шаг 1. Найдем максимум первого слагаемого $\sum_{j=1}^{n}(\mathbf{x}_j,\mathbf{f}_1)^2$ в выражении (7.13) при условии $(\mathbf{f}_1,\mathbf{f}_1)=N$.

Шаг 2. Найдем максимум второго слагаемого при условии $(\mathbf{f}_2,\mathbf{f}_2)=N$ и $(\mathbf{f}_l,\mathbf{f}_2)=0$.

Шаг 3. Найдем максимум третьего слагаемого при условии $(\mathbf{f}_3,\mathbf{f}_3)=N$, $(\mathbf{f}_l,\mathbf{f}_3)=0$, $(\mathbf{f}_2,\mathbf{f}_3)=0$.

И так далее для всех компонент.

Рассмотрим схему вычисления первой компоненты. Задача решается методом множителей Лагранжа, который основывается на следующем утверждении.

Для локального экстремума действительной функции $F(\mathbf{x}_1,...,\mathbf{x}_n)$ при m ограничениях вида $\varphi_1(\mathbf{x}_1,...,\mathbf{x}_n)=0$, ... , $\varphi_m(\mathbf{x}_1,...,\mathbf{x}_n)=0$ существует необходимое условие

$$\frac{\partial \Phi}{\partial x_1} = \frac{\partial \Phi}{\partial x_2} =...= \frac{\partial \Phi}{\partial x_n} = 0 \ ,$$

где $\Phi(\mathbf{x}_1,...,\mathbf{x}_n)= F(\mathbf{x}_1,...,\mathbf{x}_n)+\sum_{k=1}^{m} \lambda_k\varphi_k(\mathbf{x}_1,...,\mathbf{x}_n)$.

Коэффициенты λ_k называются множителями Лагранжа, а $\Phi(\mathbf{x}_1,...,\mathbf{x}_n)$ - функцией Лагранжа.

То есть задача отыскания локального экстремума $F(\mathbf{x}_1,...,\mathbf{x}_n)$ при заданном наборе ограничений сводится к отысканию глобального экстремума функции Лагранжа.

В нашей задаче функция Лагранжа имеет вид:

$$\Phi(\mathbf{f}_1, \lambda_1)=\sum_{j=1}^{n} (\mathbf{x}_j,\mathbf{f}_1)^2- \lambda_1(\mathbf{f}_1, \mathbf{f}_1). \quad (7.14)$$

Дифференцируя эту функцию по $\mathbf{f}_1$, получаем N уравнений вида:

$$\frac{\partial \Phi}{\partial f_{1i}} = 2\sum_{j=1}^{n} (\mathbf{x}_j,\mathbf{f}_1)x_{ij}-2\lambda_1 f_{i1}=0. \quad (7.15)$$

Или в векторной форме:

$$\sum_{j=1}^{n} (\mathbf{x}_j,\mathbf{f}_1)\, \mathbf{x}_j =\lambda_1\mathbf{f}_1. \quad (7.16)$$

В этой системе имеется n+m неизвестных. Для того, чтобы ее решить, используем следствие 2 из выражения (7.13) и перепишем систему в виде:

$$\sum_{j=1}^{n} \alpha_{1j}\mathbf{x}_j= \frac{1}{N} \lambda_1\mathbf{f}_1. \quad (7.17)$$

Путем последовательного домножения (7.17) на каждый из исходных параметров $\mathbf{x}_s$, s=1,...,n, эта система приводится к виду $\frac{1}{N}\sum_{j=1}^{n} (\mathbf{x}_j, \mathbf{x}_s)= \frac{1}{N} \lambda_1\mathbf{f}_1$, что эквивалентно выражению $\sum_{j=1}^{n} \rho(\mathbf{x}_j, \mathbf{x}_s)= \frac{1}{N} \lambda_1\alpha_{1s}$ или, в векторном виде,

$$\mathbf{R}\mathbf{a}_1= \frac{1}{N} \lambda_1\mathbf{a}_1. \quad (7.18)$$

Здесь **R** - корреляционная матрица исходных параметров $\mathbf{x}_j$, j=1,...,n, $\mathbf{a}_1$- первый собственный вектор матрицы **R**, а $\frac{1}{N} \lambda_1$- собственное число этой матрицы, соответствующее собственному вектору $\mathbf{a}_1$.

Домножив уравнение (7.18) на $\mathbf{f}_1$ и используя условие 2 из (7.10), можно получить условие нормировки собственного вектора $\mathbf{a}_1$:

$\lambda_1 = N\sum_{j=1}^{n} \alpha_{1j}^2$ или, иначе, $\lambda_1 = \sum_{j=1}^{n} (\mathbf{x}_j, \mathbf{f}_1)$. (7.19)

Поскольку правая часть уравнения (7.19) должна достигать максимума по условию (7.15), $\frac{1}{N}\lambda_1$ соответствует максимальному собственному числу матрицы **R**.

Таким образом, решение задачи в модели главных компонент сводится к нахождению собственных векторов и собственных чисел корреляционной матрицы **R**, ранжированных в порядке убывания собственных значений. Собственные числа матрицы **R** находятся из решения алгебраического уравнения n-й степени

$$\det(\mathbf{R}-\lambda\mathbf{I})=0, \qquad (7.20)$$

где **I** – единичная матрица размерности n x n, а направления собственных векторов – из решения системы линейных уравнений (7.18). В тех случаях, когда нам необходим только один общий фактор, например, при картографической генерализации количественных показателей в ГИС, можно воспользоваться простым итеративным методом нахождения первого собственного вектора и соответствующего ему собственного числа, описанным в [5].

В новой системе координат, где признаки независимы, корреляционная матрица становится единичной.

Если мы вспомним про взаимосвязь матрицы **R** и ковариационной матрицы **C**, элементы которой определяются в (7.1), а также что представляют собой координаты вектора $\mathbf{x}_j$ в новом базисе, то из выражения (7.19) становится ясно, что λ_1 соответствует направлению наибольшего разброса точек в пространстве **X**; следующие собственные значения будут соответствовать собственным векторам в порядке убывания разброса. Практически переход к главным компонентам чаще осуществляется по следующей схеме: ищутся все n корней уравнения (7.20) и затем выбирается требуемое количество первых m компонент в порядке убывания их значений. Эта же схема применяется при анализе главных компонент в пакетах обработки данных ДЗ. Тем не менее, если на изображении присутствует достаточно большое количестве классов различного происхождения, сильно отличающихся по своим свойствам, это преобразование может не дать желаемых результатов. При обработке видеоданных ДЗ преобразование к главным компонентам целесообразно для изображений, множество образов которых в признаковом пространстве вытянуто вдоль некоторого направления, что свидетельствует о сильной корреляции признаков. В этом случае преобразование к главным компонентам иногда позволяет извлечь дополнительную, важную для интерпретатора информацию. В тех же

случаях, когда в целом по изображению корреляция между растровыми слоями невелика, различив одни классы, мы можем «потерять» другие. В этом смысле предварительная неконтролируемая классификация более эффективна. Тем не менее, при классификации с обучением для выбора эталонных объектов и формирования обучающих выборок по классам такой подход может оказаться целесообразным, особенно для специалистов-предметников, занимающихся анализом конкретных процессов и явлений.

В общем случае, безотносительно к характеру приложения, преобразование пространства признаков с целью снижения размерности задачи целесообразно проводить только после формирования перечня выделяемых классов. Направления главных компонент существенно зависят от расположения классов в пространстве **X** и характера внутригруппового и межгруппового разброса точек множества. Если интересующие нас классы не коррелированы, расположены на множестве образов неравномерно и их среднеквадратичные отклонения сильно отличаются, предварительное преобразование признакового пространства к главным компонентам может привести в конечном итоге к значительным ошибкам. Поэтому в большинстве случаев задачу снижения размерности (отбора признаков) решают с учетом особенностей данных в конкретной задаче классификации [1, 19]. В [19], например, выделены шесть принципов снижения размерности и описаны подходы, которые могут применяться в каждом отдельном случае.

Контрольные вопросы к разделу 7.

1. Перечислите основные задачи корреляционного анализа при классификации многомерных данных.
2. Дайте определение ковариационной и корреляционной матрицы. Чем они отличаются?
3. С какой целью осуществляется переход к стандартизованной (нормализованной) матрице данных? Какими свойствами обладает эта матрица?
4. Приведите примеры задач, в которых применяется факторный анализ.
5. Какими свойствами обладают общие факторы в модели главных компонент? Что такое характерный фактор?
6. Какой показатель минимизируется (максимизируется) в задаче главных компонент?
7. Перечислите основные этапы решения задачи главных компонент в общем виде? Как эта задача решается на практике?
8. Дайте графическую интерпретацию главных компонент для признакового пространства в задаче классификации многомерных данных.
9. Что может дать анализ главных компонент при тематической обработке многозональной аэрокосмической информации? Почему этот метод отнесен в пакете ERDAS Imagine к блоку «Интерпретатор»?

8.Статистический подход к задаче распознавания. Элементы теории статистических решений.

Методы классификации, которые мы рассматривали ранее, иначе называются **классификацией по минимуму расстояния**. К этой группе методов относится не только кластерный анализ, но и ряд других процедур, в которых мерой сходства между образами является расстояние в пространстве признаков. В таком подходе предполагается, что ошибка классификации связана только со взаимным положением образа и класса в этом пространстве. Тем не менее, часто встречаются ситуации, когда один и тот же образ может принадлежать сразу нескольким классам, но частота появления образа с конкретным значением для различных классов не одинакова и определить оптимальную границу между классами можно только с учетом этого факта. При обработке изображений относительную частоту появления образа в классе можно оценить, построив гистограмму по объектам, принадлежащим заданному классу. При большом количестве образов в k-м классе эта частота стремится к вероятности $p_k(\mathbf{x})$ появления образа **x** в данном классе, и гистограмму можно рассматривать как дискретную аппроксимацию функции плотности распределения $f_k(\mathbf{x})$.

Очередной класс алгоритмов, который мы рассмотрим, широко используется в пакетах обработки данных ДЗ и некоторых других прикладных направлениях. Он объединен общим названием - **принцип максимума правдоподобия**. Эти алгоритмы основаны на статистическом подходе к задаче распознавания и представляют собой **классификаторы с обучением**. То есть предполагается, что

1) заданы определенные классы и по ним имеются данные предварительных исследований (тестовые участки при обработке данных ДЗ);
2) эти данные позволяют получить представительные выборки образов по всем классам и оценить вероятность появления образа в каждом классе для всего множества X.

Рассмотрим правила принятия решений в задаче статистической классификации на примере двух классов и одного признака, заданного действительным значением измерения X. Задача заключается в определении на шкале X интервалов Ω_1 и Ω_2, на которых будут приниматься решения в пользу первого и второго класса соответственно. Для простоты в дальнейшем классы и соответствующие им области принятия решения будем обозначать одним и тем же символом Ω.

Предположим, что получена вся необходимая статистическая информация о классах: функции плотности статистического распределения

$f_1(x)$ и $f_2(x)$ (рис.8.1), а так же априорные вероятности $P(\Omega_1)$ и $P(\Omega_2)$ появления данных классов. Для задачи обработки данных ДЗ априорные вероятности - это ожидаемые доли площади под каждым классом, которые могут быть получены, например, из фондовых материалов.

Пусть мы выбрали некоторую точку x_0, разделяющую всю шкалу значений признака x на два интервала: $(-\infty,x_0]$ соответствует области решений в пользу класса Ω_1, (x_0,∞) - области решений в пользу класса Ω_2. В теории статистических решений непротиворечивое множество предположений о свойствах случайной величины ξ называется **статистической гипотезой**. В данном случае мы рассматриваем две

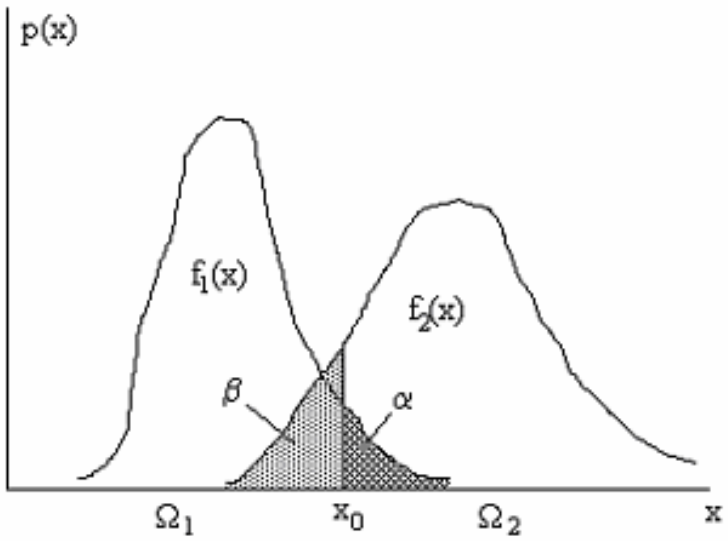

Рис.8.1. Статистическая гипотеза для случая двух классов.

статистические гипотезы: ξ, принимающая значения x, имеет распределение с плотностью $f_1(x)$, то есть принадлежит классу Ω_1, - это гипотеза H_1 против альтернативы H_2: ξ имеет распределение с плотностью $f_2(x)$, то есть принадлежит классу Ω_2.

Вероятность появления любого значения x отлична от нуля и для первого, и для второго класса на всем множестве X, поэтому при принятии решения относительно принадлежности некоторого значения x к одному из классов могут возникнуть 4 ситуации.

1. Принимаем гипотезу H_1, и она верна.
2. Принимаем H_2, но верна H_1 .
3. Принимаем H_2, и она верна.
4. Принимаем H_1, но верна H_2.

Вероятность возникновения ситуации 1 соответствует площади под $f_1(x)$ на полуинтервале $(-\infty,x_0]$, ситуации 2 - площади под $f_1(x)$ на интервале (x_0,∞), ситуации 3 - площади под $f_2(x)$ на (x_0,∞), ситуации 4 - площади под $f_2(x)$ на

$(-\infty, x_0]$. Суммарная площадь под $f_1(x)$ и $f_2(x)$ для ситуаций 2 и 4 - это **полная вероятность ошибок** в нашей схеме принятия решений. В случае двух альтернативных гипотез ошибку, соответствующую ситуации 2, обычно называют **ошибкой первого рода** (α), ошибку, соответствующую ситуации 4, - **ошибкой второго рода** (β). Вообще говоря, понятие ошибок первого и второго рода симметрично и зависит от того, какая гипотеза является основной, а какая – альтернативной. Если бы H_2 была основной гипотезой, ошибка первого рода соответствовала бы ситуации 4.

При классификации пикселов многозональных аэрокосмических изображений ошибка первого рода проявляется в появлении на объектах класса Ω_1 точек посторонних классов. В свою очередь, ошибки второго рода проявляются в появлении точек этого класса на других объектах. Когда количество классов невелико, обычно преобладают ошибки второго рода. Это связано с тем, что аналитик не учитывает все присутствующие на изображении типы объектов, в том числе и с характеристиками, близкими к выделяемым классам. Это одна из причин, по которой полезно выполнение предварительной неконтролируемой классификации, причем на значительно большее, чем требуется, количество классов. Неконтролируемая классификация позволяет предварительно оценить величину ошибок второго рода, более точно определить границы искомых классов и, при необходимости, разумно задать класс отказов от распознавания.

8.1. Байесовская стратегия минимального среднего риска.

Чтобы построить классификатор на основе описанной схемы, мы должны выработать правило оптимального разбиения всего множества X на области принятия решения в пользу каждого класса. В нашем конкретном одномерном случае - правило выбора разделяющей точки x_0. Поскольку различие между классами связано с частотой появления тех или иных значений x в этих классах, естественно выбрать точку x_0 таким образом, чтобы при многократном предъявлении классификатору образов x усредненная по совокупности решений ошибка была минимальной.

С этой целью введем “стоимость” каждого принятого решения. Для описанных выше четырех ситуаций с классами Ω_1 и Ω_2 запишем соответствующие платежные коэффициенты в виде матрицы

$$\begin{Vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{Vmatrix} . \qquad (8.1)$$

Диагональные элементы такой платежной матрицы соответствуют случаям правильной классификации (ситуации 1 и 3), два других - ошибкам (ситуации 2 и 4). Платежные коэффициенты можно задать так, чтобы они отражали наши предпочтения в отношении той или иной ситуации. Чаще

всего ограничиваются условием $c_{11}=c_{22}=0$ и $c_{12}=c_{21}=1$, то есть рассматривают только потери от ошибок.

Рассматривая платежные коэффициенты как наш "риск" в каждой из возможных ситуаций (которые можно рассматривать как случайные события), введем понятие среднего риска для описанных четырех случаев:

$$R=\sum_i r_i p(r_i),\ i=1,...,4. \quad (8.2)$$

Здесь r_i - это платежные коэффициенты из (8.1), а $p(r_i)$ - вероятность каждого из четырех событий (вероятность соответствующей выплаты). В случае, когда $c_{11}=c_{22}=0$ и $c_{12}=c_{21}=1$ (то есть «плата» - это штраф за ошибки), функцию R называют также **функцией потерь.**

Рассмотрим, что представляют собой вероятности $p(r_i)$. Например, вероятность выплаты c_{11} есть вероятность одновременного осуществления двух событий: ξ принадлежит к $f_1(x)$ (вероятность этого события в Ω_1 есть $\int_{-\infty}^{x_0} f_1(x)dx$) и появления самого класса Ω_1 (априорная вероятность этого события - $P(\Omega_1)$). То есть $p(r_1)=p(c_{11})= P(\Omega_1) \int_{-\infty}^{x_0} f_1(x)dx$.

Следовательно, формулу (8.2) можно записать так:

$$R=c_{11}\ P(\Omega_1)\int_{-\infty}^{x_0} f_1(x)dx + c_{12}\ P(\Omega_1)\int_{x_0}^{\infty} f_1(x)dx + c_{22}\ P(\Omega_2)\int_{x_0}^{\infty} f_2(x)dx + c_{21}\ P(\Omega_2)\int_{-\infty}^{x_0} f_2(x)dx. \quad (8.3)$$

Выражение (8.3) есть полный средний риск так называемой **стратегии Байеса**. Минимум R в точке x_0 достигается при условии $\left.\frac{dR}{dx}\right|_{x=x_0}=0$. Возьмем производную в точке x_0, учитывая, что $\int f(x)dx=F(x)$, $F(-\infty)=0$, $F(\infty)=1$:

$$\left.\frac{dR}{dx}\right|_{x=x_0} = c_{11}\ P(\Omega_1)\ f_1(x) - c_{12}\ P(\Omega_1)\ f_1(x) - c_{22}\ P(\Omega_2)\ f_2(x) + c_{21}\ P(\Omega_2)\ f_2(x) = 0. \quad (8.4)$$

Отсюда имеем следующее соотношение для $x=x_0$:

$$\frac{P(\Omega_1)f_1(x)}{P(\Omega_2)f_2(x)} = \frac{c_{21}-c_{22}}{c_{12}-c_{11}} = \lambda. \quad (8.5)$$

Отношение (8.5) называется **отношением правдоподобия**, а величина λ - **коэффициентом правдоподобия**. При значениях $\frac{P(\Omega_1)f_1(x)}{P(\Omega_2)f_2(x)} \geq \lambda$ решение принимается в пользу Ω_1, при $\frac{P(\Omega_1)f_1(x)}{P(\Omega_2)f_2(x)} < \lambda$ - в пользу Ω_2.

Если положить, что $c_{11}=c_{22}=0$ и $c_{12}=c_{21}=1$, получим:

$$\frac{P(\Omega_1)f_1(x)}{P(\Omega_2)f_2(x)}=1 \text{ или, в логарифмической форме, } \ln\frac{P(\Omega_1)f_1(x)}{P(\Omega_2)f_2(x)}=0. \quad (8.6)$$

Если значения признака для обоих классов распределены по нормальному закону

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad (8.7)$$

со средними m1, m2 и среднеквадратическими отклонениями σ1, σ2 соответственно, отношение правдоподобия в логарифмической форме имеет вид:

$$\ln\frac{P(\Omega_1)}{P(\Omega_2)}+\ln\frac{\sigma_2}{\sigma_1}-\frac{(x-m_1)^2}{2\sigma_1^2}+\frac{(x-m_2)^2}{2\sigma_2^2}=0. \quad (8.8)$$

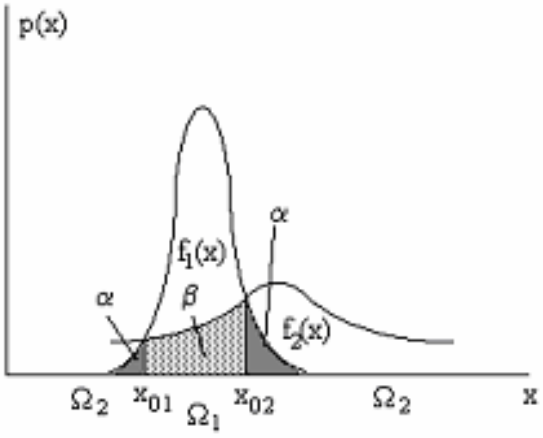

Рис.8.2. Разделение классов с разными дисперсиями признака.

То есть x_0 является решением квадратного уравнения (8.8). Случай, когда уравнение имеет два действительных корня, представлен на рисунке (8.2).

8.2. Байесовская стратегия в условиях неполной информации.

Рассмотрим еще два случая: 1) неизвестны априорные вероятности, но заданы платежные коэффициенты и функции плотности распределения; 2) известны только функции плотности распределения.

Стратегия принятия решений для двух гипотез при отсутствии априорных вероятностей называется минимаксной стратегией. В этом случае

предполагается, что $P(\Omega_2)=1- P(\Omega_1)$. Введем обозначения $P(\Omega_1)=P$, $P(\Omega_2)=1-P$ и рассмотрим выражение (8.3) как функцию P. При каком-то значении P минимальный байесовский риск $R_{min}(P)$ достигает максимума. Положим $c_{11}=c_{22}=0$ в (12.3) и найдем условие этого максимума, продифференцировав R(P) по P и приравняв производную нулю:

$$\frac{dR(P)}{dP}=c_{12}\int_{x_0}^{\infty} f_1(x)dx - c_{21}\int_{-\infty}^{x_0} f_2(x)dx = c_{12}\alpha - c_{21}\beta=0. \quad (8.9)$$

Отсюда получаем соотношение:

$$\frac{\alpha}{\beta}=\frac{c_{21}}{c_{12}},$$

в случае $c_{12}=c_{21}=1$ имеем просто $\alpha=\beta$. То есть при отсутствии априорных вероятностей наилучшая стратегия заключается в уравнивании ошибок первого и второго рода.

При отсутствии априорных вероятностей и платежных коэффициентов обычно фиксируется некоторое значение α и при этом значении ищутся условия минимума ошибки β (или максимума величины $1-\beta$, которая называется мощностью статистического критерия). Такая стратегия называется стратегией Неймана-Пирсона, а статистические критерии, основанные на ее использовании – односторонними. Односторонние критерии часто применяются при так называемых непараметрических оценках, где решение принимается непосредственно по статистической выборке без каких-либо предположений о типе статистического распределения [16].

8.3. Случай нескольких классов. Априорные и апостериорные вероятности. Формула Байеса.

Пусть теперь задано K классов $\Omega_1,...,\Omega_K$ с функциями плотности распределения признака x $f_1(x),...,f_K(x)$ и априорными вероятностями $P(\Omega_1),...,P(\Omega_K)$. К примеру, мы хотим разбить на K классов точки серого полутонового изображения по значениям яркости x. Гистограмма такого изображения, нормированная к общему количеству точек S_0, представляет собой дискретную аппроксимацию плотности p(x) совместного распределения точек K классов. Рассмотрим, что представляет из себя значение p(x) в отдельно взятой точке x_i.

Как мы уже отметили при записи среднего байесовского риска, вероятность появления точки со значением x_i в k-м классе есть произведение вероятностей $P(\Omega_k)$ $f_k(x_i)$, или, иначе, $P(\Omega_k)p(x_i/\Omega_k)$. Плотность совместного распределения в точке x_i - полная вероятность появления значения x_i - есть $\sum_{j=1}^{K} P(\Omega_j)$ $f_j(x_i)$. Общую долю k-го класса в этой сумме, то есть вероятность

того, что точка со значением x_i принадлежит именно k-му классу, тогда можно записать так

$$p(\Omega_k / x_i) = \frac{P(\Omega_k) p(x_i / \Omega_k)}{\sum_{j=1}^{K} P(\Omega_j) p(x_i / \Omega_j)}. \quad (8.10)$$

В теории вероятностей формулу (8.10) называют **формулой Байеса** для полной системы несовместных событий. В нашем случае такую систему образуют решения об отнесении точки со значением x_i к одному из К классов.

В отличие от априорной вероятности $P(\Omega_k)$, которая характеризует ожидаемую частоту появления k-го класса по множеству образов в целом, $p(\Omega_k/x_i)$ - это вероятность появления k-го класса при конкретном значении признака x. Эту вероятность называют **апостериорной вероятностью**.

Поскольку $P(\Omega_k)p(x/\Omega_k)=P(\Omega_k)f_k(x)$ (величину $P(\Omega_k)p(x/\Omega_k)$ часто называют **функцией правдоподобия**[*]), а знаменатели в выражении (8.10) одинаковы для всех К классов, мы можем для любой пары классов k и s записать отношение правдоподобия (8.6) через апостериорные вероятности

$$L_{ks}(x) = \frac{p(\Omega_k / x)}{p(\Omega_s / x)}. \quad (8.11)$$

В соответствии с байесовской стратегией, решение принимается в пользу того класса, для которого $L_{ks}(x)>\lambda$ ($L_{ks}(x)>1$ при $c_{12}=c_{21}=1$; в принципе, мы всегда можем выполнить соответствующую нормировку, чтобы это условие выполнялось. Следовательно, образ с конкретным значением признака x мы должны отнести к тому из К классов, для которого $p(\Omega_k /x)> p(\Omega_s /x)$ или, что эквивалентно, $P(\Omega_k)p(x/\Omega_k)>P(\Omega_s)p(x/\Omega_s)$) по всем парам k,s. Отсюда и название данного подхода - **принцип максимума правдоподобия**.

Часто при классификации по максимуму правдоподобия априорные вероятности появления всех классов считаются одинаковыми, а платежные коэффициенты принимаются равными 0 и 1, как мы это делали ранее. Классификацию при таких условиях называют **нуль-единичной байесовской стратегией**. В этом случае мы сокращаем ошибку первого рода для маловероятных классов, что позволяет выделять их более надежно.

---

[*] В некоторых работах функцией правдоподобия называют апостериорную вероятность (8.10) [2,30], а в некоторых – саму плотность распределения $p(x/\Omega_k)$. Последнее справедливо только в том случае, если для всех классов $P(\Omega_k)=1$. В любом случае под функцией правдоподобия подразумевается решающая функция, определенная ранее в разделе 5.

Тем не менее, для ряда задач априорные вероятности могут иметь важное значение. Например, при дистанционном мониторинге некоторой территории, на которой наблюдается определенное количество классов с известными $p(x/\Omega_s)$. В этом случае в результате итеративного процесса классификации с априорными вероятностями $P_i(\Omega_s)=S_{i-1}(\Omega_s)$, где i – очередной шаг итерации, а $S_{i-1}(\Omega_s)$ – доля площади под классом Ω_s, полученная на i-1-м шаге, можно получить наиболее близкие к истинным значения площадей под классами [28].

8.4. Байесовский классификатор для многомерного случая.

Все полученные выше результаты справедливы и когда образ представлен не значением одного параметра, а n-мерным вектором **x**. В этом случае логарифм отношения правдоподобия (8.6) для пары классов k,s задает уравнение разделяющей функции (5.1) $d_{ks}(\mathbf{x})=L_{ks}(\mathbf{x})=0$ в пространстве **X**. Ясно, что в общем случае это может быть гиперповерхность совершенно произвольной формы. В подобном случае использование многомерного байесовского классификатора затрудняется сложностью оценки плотностей распределения вероятностей $p(\mathbf{x}/\Omega_k)= f_k(\mathbf{x})$. Если **x** принимает ограниченное число значений, эти оценки вполне реальны. Но для задач классификации изображений, когда каждый признак (координата вектора **x**) может принимать значения в диапазоне [0,255], а количество тестовых данных для обучения ограничено, решение задачи для произвольных $f_k(\mathbf{x})$ представляет серьезные трудности. Поэтому обычно ограничиваются гипотезой о конкретном типе распределения значений признаков. Чаще всего используется гипотеза о **многомерном нормальном распределении**. В самом деле, если флуктуации яркости внутри каждого класса Ω_k обусловлены очень большим количеством различных факторов, можно считать, что яркостные характеристики класса распределены по нормальному закону $N_k(\mathbf{m}_k,\sigma_k)$. Плотность многомерного нормального распределения имеет вид

$$f_k(\mathbf{x})=p(\mathbf{x}/\Omega_k)=1/[(2\pi)^{n/2}|\mathbf{C}_k|^{1/2}]\exp[-1/2(\mathbf{x}-\mathbf{m}_k)^T\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)]. \quad (8.12)$$

Здесь $\mathbf{C}_k$ - ковариационная матрица $\mathbf{C}_k=\{\sigma_{ij}^k\}=M[(x_i-m_i^k)(x_j-m_j^k)]$, i=1,...,n, j=1,...,n, k=1,...,S_k, S_k - количество образов в доступной выборке по k-му классу, M - выборочное среднее (см. 8.1).

Выражение $(\mathbf{x}-\mathbf{m}_k)^T\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)$ называют **квадратичным расстоянием Махаланобиса** между точкой множества Ω_k и его центром $\mathbf{m}_k$. В случае, когда все n признаков независимы (т.е. образуют ортогональный базис в пространстве **X**), матрица $\mathbf{C}_k$ диагональна и расстояние Махаланобиса приобретает достаточно простой вид:

$$(\mathbf{x}-\mathbf{m}_k)^T\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)=\sum_{j=1}^{n}\frac{(x_j-m_{jk})^2}{\sigma_j^2}. \quad (8.13)$$

Решающую функцию (5.2) для k-го класса в многомерном случае удобно записывать в логарифмической форме:

$$r_k(\mathbf{x})=\ln P(\Omega_k)+\ln p(\mathbf{x}/\Omega_k)=\ln P(\Omega_k)+(1/2)\ln|\mathbf{C}_k|-(1/2)(\mathbf{x}-\mathbf{m}_k)^T\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k). \quad (8.14)$$

Разделяющая функция $d_{ks}(\mathbf{x})=0$ для классов k и s, выраженная через отношение правдоподобия (8.6) в логарифмической форме, выглядит так:

$$\ln\frac{P(\Omega_k)}{P(\Omega_s)}+(1/2)\ln\frac{|C_s|}{|C_k|}-(1/2)(\mathbf{x}-\mathbf{m}_k)^T\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)+(1/2)(\mathbf{x}-\mathbf{m}_s)^T\mathbf{C}_s^{-1}(\mathbf{x}-\mathbf{m}_s)=0. \quad (8.15)$$

В общем случае такие разделяющие функции в n-мерном пространстве X могут представлять собой гиперповерхности сложной формы - так называемые **гиперквадрики**. Более просто обстоит дело в случае, когда ковариационные матрицы одинаковы для всех K классов: $\mathbf{C}_k=\mathbf{C}_s=\mathbf{C}$. Уравнение (8.15) тогда принимает следующий вид:

$$d_{ks}(\mathbf{x})=\ln\frac{P(\Omega_k)}{P(\Omega_s)}-(1/2)(\mathbf{x}-\mathbf{m}_k)^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_k)+(1/2)(\mathbf{x}-\mathbf{m}_s)^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}_s)=$$

$$\ln\frac{P(\Omega_k)}{P(\Omega_s)}+\mathbf{x}^T\mathbf{C}^{-1}(\mathbf{m}_k-\mathbf{m}_s)-(1/2)(\mathbf{m}_k+\mathbf{m}_s)^T\mathbf{C}^{-1}(\mathbf{m}_k-\mathbf{m}_s)=0. \quad (8.16)$$

То есть при $\mathbf{C}_k=\mathbf{C}_s=\mathbf{C}$, $\mathbf{m}_k\neq\mathbf{m}_s$ гиперквадрики превращаются в линейные разделяющие функции. Заметим, что в этом случае использование классификатора по максимуму правдоподобия не дает никакого преимущества перед классификатором по минимуму расстояния, если мы работаем с нуль-единичной стратегией, то есть считаем равными априорные вероятности появления классов. Поэтому если мы заранее установили, что дисперсии классов приблизительно одинаковы, можно использовать более простой метод классификации практически безо всякого ущерба для конечного результата.

В случае, когда $\mathbf{m}_k=\mathbf{m}_s,=\mathbf{m}$, $\mathbf{C}_k\neq\mathbf{C}_s$, функция (8.15) приобретает вид:

$$\ln\frac{P(\Omega_k)}{P(\Omega_s)}+(1/2)\ln\frac{|C_s|}{|C_k|}-(1/2)(\mathbf{x}-\mathbf{m})^T(\mathbf{C}_k^{-1}-\mathbf{C}_s^{-1})(\mathbf{x}-\mathbf{m})=0. \quad (8.17)$$

Таким образом, можно сказать, что практическим преимуществом многомерного байесовского классификатора перед классификаторами по минимуму расстояния является возможность разделения классов, имеюших равные средние, но разный разброс признаков. Тем не менее, это преимущество в явном виде используется достаточно редко. Однако уже сама возможность учета дисперсии признаков дает возможность лучше разделять классы в тех случаях, когда разброс признаков в классах достаточно велик в сравнении с полным диапазоном их значений (в обработке многозональных изображений, например, при классификации лесов или селитебных зон). Основной проблемой, как уже говорилось, остается получение репрезентативных данных для построения функций правдоподобия. Поэтому главной задачей аналитика данных при использовании такого классификатора является тщательный подбор эталонов классов (обучающих выборок) с предварительным исследованием всех факторов, которые могут повлиять на результат классификации.

8.5. Обучение статистических классификаторов. Вероятность ошибок. Меры статистической разделимости.

Обучение классификаторов – задача достаточно сложная, и ее особенности существенно зависят от сферы приложения алгоритма. В общем случае необходимо проведение целого ряда исследований: прежде всего, выбор гипотезы о типе статистического распределения используемых признаков в каждом из классов и затем проверка этой гипотезы на репрезентативных выборках по объектам исследования. В тематической обработке многозональных аэрокосмических изображений это один из самых проблематичных этапов. Помимо ограничений конкретного алгоритма классификации, результат, который мы получаем на выходе, зависит еще и от следующих факторов.

1. От способа синтеза изображения при выводе на экран и особенностей палитры. В процессе синтеза так или иначе происходит некоторая потеря информации, обусловленная самой математической моделью процесса. Кроме того, индивидуальные особенности восприятия человеком цветовой гаммы влияют на способность аналитика различать объекты исследования и, следовательно, на процесс формирования обучающих выборок.
2. От достоверности наземных данных, в том числе от способа их сбора и точности координатной привязки тестовых участков к изображению. Даже при разрешении цифрового на местности 30-50м мы едва ли сможем строго привязать точечные измерения. Следовательно, для надежного обучения классификатора необходимы тестовые участки с

высокой степенью пространственной однородности. При работе с материалами наземных обследований это еще один довод в пользу двухэтапной схемы обработки: сначала выполнение неконтролируемой классификации с целью выбора однородных участков для получения наземных данных, потом проведение наземных обследований, затем классификация с обучением. В настоящее время, однако, такой схемой пользуются редко из-за высокой стоимости детальных наземных обследований и ищут другие пути решения этой проблемы.

3. Наконец, от количества и способа расположения обучающих выборок на изображении. Изменчивость спектральных отражательных свойств объектов земной поверхности даже в пределах одного изображения может оказаться очень высокой. Кроме факторов, обусловленных условиями съемки и учитывающихся в процессе нормализации изображений, на отражательные характеристики объектов могут повлиять такие факторы, как ветер, осадки и т.п.

Предположим, тем не менее, что мы получили вполне надежные обучающие выборки и рассчитали параметры функций плотности распределения для всех классов. Каким способом можно оценить возможности удовлетворительной классификации конкретных данных при выбранных описаниях классов, то есть ожидаемые вероятности ошибок? Для этой цели используется такое понятие, как статистическая разделимость.

Статистическая разделимость классов - это некоторая функциональная характеристика, известным образом связанная с вероятностью ошибки классификации.

Как мы уже видели при рассмотрении статистического классификатора, вероятность ошибки при разделении пары классов связана с площадью перекрытия их функций плотности распределения.

Ясно, что при различных средних значениях m_1 и m_2 (рис. 8.3, а) вероятность ошибки убывает при увеличении расстояния между средними. Поэтому для классов с различными средними можно использовать такую меру статистической разделимости, как нормализованное расстояние:

$$R_n = \frac{|m_1 - m_2|}{\sigma_1 + \sigma_2}. \quad (8.18)$$

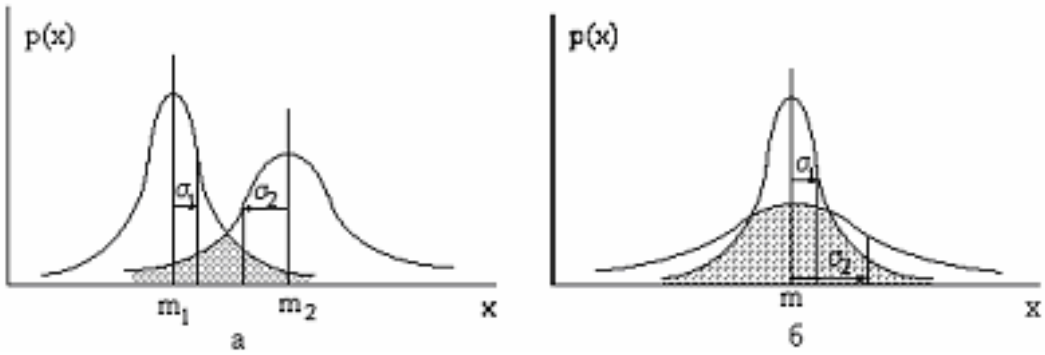

Рис.8.3. Ошибки классификации при разных средних значениях признака (а) и одинаковых средних (б).

Величина R_n связана обратной зависимостью с вероятностью ошибок: она возрастает как с увеличением расстояния между средними, так и с уменьшением дисперсии внутри классов. Для нормальных распределений эта величина принимает значение $R_n=1$ “на уровне одного σ”, то есть когда функции плотности распределения соприкасаются в точках перегиба: $|m_1-m_2|=\sigma_1+\sigma_2$. На этом уровне при равных априорных вероятностях появления классов вероятность ошибок классификации (заштрихованная площадь) $\varepsilon\approx0.33$. Это обычно считается верхней границей допустимой величины ошибки, то есть результат уже поддается интерпретации и постклассификационной обработке.

Для многомерного случая (при $\mathbf{C}_1=\mathbf{C}_2=\mathbf{C}$, $\mathbf{m}_1\neq\mathbf{m}_2$) иногда используется квадратичное расстояние Махаланобиса между векторами средних значений, которое для нормальных распределений иначе называют расстоянием между плотностями распределения:

$$R_{12}=(\mathbf{m}_1-\mathbf{m}_2)^{T}\mathbf{C}^{-1}(\mathbf{m}_1-\mathbf{m}_2). \quad (8.19)$$

Расстояние (8.19) также обратно пропорционально ожидаемой величине ошибки.

Недостатки мер такого типа проявляются в случаях, когда средние значения для двух классов совпадают (рис.8.3, б). Поэтому для обработки всех ситуаций необходима мера более универсальная. Таковой является например, попарная **дивергенция**, позволяющая учесть соотношения между значениями плотностей распределения двух классов в каждой точке x.

Дивергенция определяется через так называемое **среднее количество различающей информации.**

Средним количеством различающей информации для класса Ω_1 относительно класса Ω_2 называется величина

$$\int_X p(x/\Omega_1)\ln\frac{p(x/\Omega_1)}{p(x/\Omega_2)}dx. \quad (8.20)$$

Аналогично можно записать среднее количество различающей информации для класса Ω_2 относительно класса Ω_1. Тогда полное среднее количество различающей информации для пары классов Ω_1 и Ω_2 будет выглядеть так:

$$D_{12} = \int_X p(x/\Omega_1)\ln\frac{p(x/\Omega_1)}{p(x/\Omega_2)}dx + \int_X p(x/\Omega_2)\ln\frac{p(x/\Omega_2)}{p(x/\Omega_1)}dx = \int_X (p(x/\Omega_1) - p(x/\Omega_2))\ln\frac{p(x/\Omega_1)}{p(x/\Omega_2)}dx. \quad (8.21)$$

Величина D_{12} называется дивергенцией. Как видно из формулы (8.21), дивергенция включает само отношение правдоподобия и его логарифм, то есть учитывает расстояние между функциями правдоподобия в каждой точке пространства **X** и их соотношение.

Несмотря на кажущуюся сложность выражения (8.21) для многомерного случая, для нормально распределенных значений **x** дивергенция достаточно просто вычисляется через средние и и ковариационные матрицы пары классов.

$$D_{12} = ?\ \mathrm{tr}\,[(\mathbf{C}_1-\mathbf{C}_2)(\mathbf{C}_2^{-1}-\mathbf{C}_1^{-1})] + ?\ \mathrm{tr}\,[(\mathbf{C}_1^{-1}+\mathbf{C}_2^{-1})(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T. \quad (8.22)$$

Дивергенция пригодна для оценки обеих ситуаций, представленных на рис. 8.3. Она удовлетворяет требованиям метрики: $D_{ij}>0$ при $i\neq j$, $D_{ij}=0$ при $i=j$, $D_{ij}=D_{ji}$, $D_{ij}(x_1, \ldots, x_n)\leq D_{ij}(x_1, \ldots, x_n, x_{n+1})$, то есть добавление нового измерения никогда не приводит к уменьшению дивергенции. Более того, если признаки классов распределены по нормальному закону с равными ковариационными матрицами, то нетрудно показать, что $D_{ij}=R_{ij}$, где R_{ij} – квадратичное расстояние Махаланобиса (8.19). При статистической независимости измерений дивергенция аддитивна: $D_{ij}(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} D_{ij}(x_k)$. Эти свойства дивергенции иногда используют при выборе признаков для оценки их информативности [1]: при заданном количестве классов из возможного набора признаков следует отбирать те, для которых общая или средняя попарная дивергенция максимальна.

Однако квадратичная мера D с увеличением расстояния между классами растет значительно быстрее, чем величина R из (8.18). Поэтому использование для набора из К классов такой оценки, как средняя попарная дивергенция, целесообразно только в тех случаях, когда все классы распределены равномерно по пространству X. В противном случае даже один класс, далеко отстоящий от всех остальных, может дать слишком оптимистическую оценку ошибки.

В таких ситуациях предпочтительнее использовать так называемое расстояние Джеффриса-Матуситы (J-M расстояние) [4,32]. J-M расстояние рассчитывается по формуле:

$$J_{12} = \{\int_X [\sqrt{p(x/\Omega_1)} - \sqrt{p(x/\Omega_2)}]^2 dx\}^{\frac{1}{2}}. \quad (8.23)$$

Для нормально распределенных значений признака поведение J-M расстояния связано с ожидаемой вероятностью правильного распознавания зависимостью, которую можно считать почти линейной.

Контрольные вопросы к разделу 8.

1. В каких задачах распознавания целесообразно применение статистического подхода? Какие данные об объектах распознавания необходимы для его реализации?
2. Объясните содержательно понятие ошибок первого и второго рода. Как проявляются ошибки первого и второго рода при классификации пикселов изображения?
3. Объясните, как влияют на значение коэффициента правдоподобия 1) априорные вероятности появления классов; 2) значения платежных коэффициентов. Что называют нуль-единичной байесовской стратегией и в каких случаях ее целесообразно использовать?
4. Какие стратегии принятия решения, основанные на байесовском подходе, используются: 1) при неизвестных априорных вероятностях; 2) при неизвестных априорных вероятностях и платежных коэффициентах?
5. С какой целью при классификации пикселов изображения широкодисперсные классы часто разбивают на несколько подклассов?
6. Что такое апостериорная вероятность? Продемонстрируйте графически принцип принятия решения по максимуму правдоподобия в одномерном случае для трех классов с различными средними и дисперсиями.
7. В каком случае результат классификации по принципу максимума правдоподобия совпадает с результатом классификации по минимуму расстояния? Объясните это аналитически.
8. Какие факторы влияют на качество обучения классификатора при обработке аэрокосмических изображений?
9. Перечислите основные способы предварительной оценки качества классификации по принципу максимума правдоподобия. В каких случаях целесообразно использовать каждый из этих способов?

9. Самообучающиеся системы распознавания.

Самообучающиеся системы распознавания отличаются от обычных систем распознавания с обучением тем, что для построения оптимальных разделяющих или решающих функций используется итеративная

автоматическая процедура обучения по заданной выборке образов. Как правило, используется линейное или обобщенное линейное представление этих функций, которое в большинстве случаев удается получить путем соответствующих преобразований. В этом случае при количестве признаков n задача обучения классификатора сводится к нахождению вектора весовых коэффициентов **w** размерности n+1, обеспечивающего справедливость системы неравенств

$$\mathbf{Xw}>0, \quad (9.1)$$

где **X** – матрица размерности n x N, N – общее количество дискриминантных функций. Если вектор **w** существует, то система неравенств (9.1) совместна. На языке распознавания образов это означает разделимость соответствующих классов.

Свободный n+1 – й член уравнения на практике имеет смысл порогового ограничения, который переносится в левую часть при аналитическом решении задачи.

Основные типы алгоритмов обучения - метод корректирующих приращений, метод минимальной среднеквадратичной ошибки, метод потенциальных функций – могут быть построены как на основе детерминистского, так и в статистического подхода. Существуют корректные доказательства сходимости алгоритмов такого типа, некоторые из которых приведены в [1]. Мы рассмотрим только общую схему обучения указанных типов алгоритмов в рамках методологии линейного дискриминантного анализа.

9.1. Алгоритм корректирующих приращений.

Этот подход иначе называют **алгоритмом перцептрона** или **«подкрепления-наказания».**

Модель перцептрона, предложенная в 1957г. Розенблаттом, в настоящее время получила свое развитие в нейронных сетях. Суть обучающего алгоритма состоит в следующем. Пусть имеется некоторое сенсорное устройство R, на вход которого поступает совокупность сигналов с n других сенсорных устройств. Устройства, подающие сигналы на вход системы-датчика R либо непосредственно регистрируют состояние внешней среды, либо являются такими же сенсорными устройствами, обрабатывающими сигналы с других датчиков. В зависимости от их значений устройство R вырабатывает собственный сигнал, величина которого зависит от его реакции f_R на состояние входа. Данные, поступающие на вход системы R, представляют собой вектор **x** размерности n. Реакция системы

пропорциональна значениям x_j, j=1,…,n, взятым с весами w_j, которые учитывают вклад того или иного сигнала:

$$f_R(\mathbf{x}) = \sum_{j=1}^{n+1} w_j x_j = \mathbf{w}^T \mathbf{x}, \quad (9.2)$$

Если значение w_{n+1} соответствует желаемой величине сигнала на выходе, взятой с обратным знаком, то правильная реакция системы будет соответствовать значению $f_R(\mathbf{x})=0$.

Процесс обучения в данном случае будет заключаться в определении вклада каждого сигнала с датчика предыдущего уровня в реакцию датчика R. Предположим, что начальный вектор весов $\mathbf{w}(1)$, соответствующий шагу 1 алгоритма обучения, выбран произвольно. Выберем некоторое корректирующее приращение c. Коррекция весов производится по правилу:

$$\mathbf{w}(k+1) = \begin{cases} \mathbf{w}(k) - c\mathbf{x}(k), & f_R(\mathbf{x}) > 0, \\ \mathbf{w}(k), & f_R(\mathbf{x}) = 0, \\ \mathbf{w}(k) + c\mathbf{x}(k), & f_R(\mathbf{x}) < 0. \end{cases} \quad (9.3)$$

С точки зрения геометрической интерпретации задачи распознавания функция $f_R(\mathbf{x})$ соответствует разделяющей функции $d_{ij}(x)$ для пары классов i,j.

Таким путем в процессе обучения отыскивается оптимальное положение гиперплоскости, разделяющей эти классы. В случае, если классы линейно разделимы, алгоритм обучения сходится за конечное число шагов.

9.2. Алгоритм минимальной среднеквадратичной ошибки (НСКО-алгоритм обучения).

В данном случае, как и в НСКО-алгоримах кластеризации, вводится функционал качества обучения

$$J(\mathbf{w},\mathbf{x},\mathbf{b}) = \sum_{i=1}^{N} (\mathbf{w}^T \mathbf{x}_i - \mathbf{b}) = \frac{1}{2} \| \mathbf{X}\mathbf{w} - \mathbf{b} \|^2, \quad (9.4)$$

где **b** – взятый с обратным знаком вектор реакций системы по всем j входам, j=1,…,n. Данный алгоритм называют иначе **алгоритмом Хо-Кашьяпа**.

Минимум функционала ищется из решения системы уравнений:

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{w}} &= \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{b}) = 0, \\ \frac{\partial J}{\partial \mathbf{b}} &= -(\mathbf{X}\mathbf{w} - \mathbf{b}) = 0. \end{aligned} \quad (9.5)$$

Анализ системы (9.5) приводит к следующим условиям существования решения:

$$\mathbf{e}(k)=\mathbf{X}\mathbf{w}(k)-\mathbf{b}(k)=0,\ \mathbf{b}(k)>0,$$

которые для задачи распознавания являются условиями разделимости классов.

Здесь k – рассматриваемый общий шаг итеративного процесса обучения.

Правила коррекции весов и вектора b, полученные из решения системы (9.5), имеют следующий вид.

$$\begin{aligned}\mathbf{w}(k+1) &= \mathbf{w}(k) + c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{b}[\mathbf{e}(k)+|\mathbf{e}(k)|],\\ \mathbf{b}(k+1) &= \mathbf{b}(k) + c[\mathbf{e}(k)+|\mathbf{e}(k)|].\end{aligned} \quad (9.6)$$

Достоинством данного алгоритма является его более высокая сходимость в сравнении с предыдущим алгоритмом обучения, а также возможность оценки разделимости классов в процессе обучения.

9.3. Метод потенциальных функций.

В данном алгоритме разделяющая два класса функция $d_{ij}(\mathbf{x})$ строится на основе последовательностей потенциальных функций $K(\mathbf{x},\mathbf{x}_1)$, $K(\mathbf{x},\mathbf{x}_2)$, …, где $\mathbf{x}_1$, $\mathbf{x}_2$, … - образы из обучающей выборки. Ясно, что метод связан с физическим понятием потенциала поля, убывающего по мере удаления от его источника. При построении системы распознавания на основе данного подхода предполагается, что объекты обучающей выборки по каждому классу создают вокруг себя некоторое потенциальное поле, то есть в каждой точке признакового пространства **X** находится единичный заряд. Тогда, по аналогии с физикой, поле, создаваемое зарядами, можно описать потенциалом, создаваемым системой зарядов во всем пространстве **X**.

В общем случае потенциальная функция задается в виде разложения

$$K(\mathbf{x},\mathbf{x}_k)=\sum_{j=1}^{\infty}\lambda_j^2\varphi_j(\mathbf{x})\varphi_j(\mathbf{x}_k). \quad (9.7)$$

При этом $\varphi(\mathbf{x})$ полагаются ортонормированными, а действительные числа λ выбираются так, чтобы функции (9.7) были ограничены.

В простейшем случае потенциальные функции могут, например, иметь следующий вид [11]:

$$K(\mathbf{x},\mathbf{x}_k)=1/(1+\alpha\,\|\,\mathbf{x}-\mathbf{x}_k\,\|) \text{ или } K(\mathbf{x},\mathbf{x}_k)=\exp(-\alpha\,\|\,\mathbf{x}-\mathbf{x}_k\,\|),$$

где α - некоторая положительная константа, $\|\mathbf{x}-\mathbf{x}_k\|$ - евклидово расстояние.

Обучающей выборке образов будет соответствовать определенная последовательность таких функций. На каждом шаге итерации алгоритма обучения строится **кумулятивный потенциал**, определяющийся по совокупности значений этих функций. Значения кумулятивного потенциала

корректируются по правилам приращения, аналогичным рассмотренным в предыдущих алгоритмах. В некоторых случаях кумулятивный потенциал может может совпадать с разделяющей функцией $d_{ij}(\mathbf{x})$. В целом же подбор вида потенциальных функций представляет собой достаточно сложную математическую задачу и требует от разработчика алгоритма достаточно большого опыта и понимания задачи.

В заключение нужно отметить, что большинство специалистов, как разработчиков, так и пользователей систем распознавания убеждены, что **распознавание образов – одновременно и наука, и искусство.** Это, в принципе, можно отнести к любому направлению, связанному с моделированием, разработкой и эксплуатацией систем искусственного интеллекта. Недаром подробное и всеобъемлющее руководство пользователя по пакету ERDAS Imagine в вопросах проектирования конкретных систем тематической обработки отправляет пользователя к опыту разработки таких систем для соответствующих прикладных направлений. Несмотря на то, что в руководстве приводятся преимущества и недостатки каждого метода классификации, все нюансы их применения могут быть выявлены только в процессе практической работы.

Вопросы применения тех или иных алгоритмов и схем обработки при тематическом дешифрировании аэрокосмических изображений предполагают не только визуальный анализ объектов исследования, но и апробацию нескольких вариантов автоматизированной обработки с обязательным анализом свойств выделяемых классов. «Кнопочное» применение тех или иных алгоритмов классификации, имеющихся в пакетах обработки, зачастую приводит к неудовлетворительным результатам и разочарованию либо в методах обработки, либо в информативности исходных данных. Любой, даже хорошо отработанный алгоритм классификации требует осмысленной настройки с пониманием сути задачи и физической природы данных, к какой бы области они не относились: будь то дистанционное зондирование, географический анализ, медицинская диагностика или социологические исследования. Нельзя забывать, что «интеллект» любой информационной системы прежде всего зависит от интеллекта ее разработчика и ее пользователя.

Литература.

Основная.

1. Дж.Ту, Р.Гонсалес. Принципы распознавания образов. М., Мир, 1978.
2. Р.Дуда, П.Харт. Распознавание образов и анализ сцен. М., Мир, 1976.
3. А.Л.Горелик, В.А.Скрипкин. Методы распознавания. М., Высшая школа, 1977.
4. Дистанционное зондирование: количественный подход. Ред. Ф.Свейн, Ш.Дэйвис. М., Недра, 1983.
5. Э.М.Браверман, И.Б.Мучник. Структурные методы обработки эмпирических данных. М., Наука, 1983.
6. Дюран Б., Одел П. Кластерный анализ. М., Статистика, 1977.

Дополнительная.

7. У.Гренандер. Лекции по теории образов. Т.1. Синтез образов. М.Мир.1979.
8. У.Гренандер. Лекции по теории образов. Т.2. Анализ образов. М.Мир.1981.
9. Классификация и кластер. Ред. Дж.Вэн Райзин. М.Мир, 1980г.
10. Прэтт У. Цифровая обработка изображений. М., Мир, 1982.
11. Адзерихо К.С., Киселевский Л.И., Костюкевич С.Б., Краснопрошин В.В.. Физические основы дистанционного зондирования. Минск, Ун-т, 1991.
12. Автоматизированная обработка изображений природных комплексов Сибири.Ред. А.Л.Яншин, В.А.Соловьев. Новосибирск, Наука,1988г.
13. Андреев Г.Г., Бочаров В.П., Сазонов Н.В., Чабан Л.Н. Автоматизация тематической обработки космических изображений при оценке состояния сельскохозяйственных культур. Исследования Земли из космоса, N5, 1986, с. 95-102.
14. Андреев Г.Г. Кравцова В.И., Михайлов В.И., Чабан Л.Н. Определение возможностей космического мониторинга необследованных территорий на основе сканерной информации высокого и среднего разрешения . Исследования Земли из космоса. 1996, N 3. - с. 57 - 67.
15. Андреев Г.Г., Афанасьева Т.В., Чабан Л.Н.. Автоматизированная тематическая обработка многозональных космических изображений пойм при составлении почвенных карт и оценке качества пойменных земель. Исследования Земли из космоса, 1997, N1. - с.
16. Андреев Г.Г., Чабан Л.Н. Методика автоматизированной тематической обработки многозональной космической информации при отсутствии или недостатке наземных данных. Исследования Земли из космоса, N 2, 1999. - с. 40-52.
17. Б.Я.Долгополов, М.Ю.Захаров, Е.А.Лупян. Комплекс программ сегментации и классификации многозональных спутниковых изображений. Исследования Земли из космоса, 1993, N6.
18. В.Б.Кашкин. А.И.Сухинин. Дистанционное зондирование Земли из космоса. Цифровая обработка изображений.,М.Логос, 2001г.
19. Патрик Э. Основы теории распознавания образов. М., Сов.радио, 1980.

20. Earl Goze. Pattern Recognition & Image Analysis. - Prentice Hall PTR. 1996.
21. Коулмен Т.Б., Эндрис Х.С.. Сегментация изображений при помощи автоматической классификации. ТИИЭР, 1979, т.67, №5, с.82-97.
22. Kruse F.A., Lefkoff F.N., Boardman J.B., Heidebrecht K.B., Shapiro A.T., Barloon P.J., Coetz A.F.H. 1993, The spectral Image Processing Systems (SIPS) – Interactive Visualization and Analysis of Imaging spectrometer Data: Remote Sesing and Environment, V. 44, p. 145-163.
23. Richards J. Remote Sensing. Digital Image Analysis. An Introduction. Second Edition. Berlin: Springer-Verlag,1993.
24. Jain A.K., Dubes R.C. Algorithms for Clastering Data. – Englewood Cliffs (NJ): Prentice-Hall, 1988.
25. M.Taraclioti, M.Petrou. Illumination Invariant Unmixing of Set of Mixed Pixels. IEEE Transaction of Geoscience & Remote Sensing, 1998, V. 39, № 10, p. 2227-2234.
26. Chain-I Chang and oth. Least Squares Subspace Projection Approach to Mixed Pixel Classification for Hiperspectral Images. - IEEE Transaction of Geoscience & Remote Sensing, 1998, V. 36, № 3, p.899-912.
27. Bea Thai, Glenn Healey. Invariant Subpixel Material Detection in Hiperspectral Imagery. - IEEE Transaction of Geoscience & Remote Sensing, 2002, V. 40, № 3, p.599-608.
28. B.Gorte, Alfred Stein. Bayesian Classification and Class Area Estimation of Satellite Imagies Using Stratification. - IEEE Transaction of Geoscience & Remote Sensing, 1998, V. 36, № 3, p.803-812.
29. Te-Ming Tu and oth. A Fast Two-Stage Classification Method for High-Dimensional Remote Sensing Data. - IEEE Transaction of Geoscience & Remote Sensing, 1998, V. 36, № 1, p.182-191.
30. C.H.M. van Kemenade, J.A.La Porte, R.J.Mokken. The ShortCut of Report: Density-Based unsupervised classification for remote sensing. Amsterdam, 1998.
31. V.D.Ripley. Pattern Recognition and Neural Networks. Cambridge University Press, 1996.
32. Tuan Tsai and William D. Philpot. A Derivative-Aided Hiperspecrtal Image Analysis System for Land-Cover Classification. - IEEE Transaction of Geoscience & Remote Sensing, 2002, V. 40, № 2, p.416-425.

Содержание.

Стр.

Введение..... 3

1. Основные понятия теории распознавания образов..... 5

2. Принципы построения систем распознавания образов..... 9

3. Особенности систем и методов распознавания в задачах тематической обработки данных ДЗ……. 12

4. Основные принципы построения систем распознавания данных ДЗ... 16

5. Задача классификации образов, представленных набором измеряемых параметров. Геометрическая и аналитическая интерпретация…….. 22

6. Группировка образов с использованием функций расстояния (кластерный анализ). Задачи кластерного анализа в тематической обработке данных ДЗ……... 26

6.1. Методы выявления (выращивания) кластеров..... 29

6.2. Кластеризация при заданном количестве групп. Алгоритмы класса ISODATA…….. 31

6.3. Дисперсионно-ковариационные критерии качества кластеризации……. 36

7. Выбор признаков. Корреляционный анализ данных..... 40

7.1. Элементы факторного анализа. Метод главных компонент..... 44

8.Статистический подход к задаче распознавания. Элементы теории статистических решений..... 50

8.1. Байесовская стратегия минимального среднего риска..... 52

8.2. Байесовская стратегия в условиях неполной информации..... 54

8.3. Случай нескольких классов. Априорные и апостериорные вероятности. Формула Байеса..... 55

8.4. Байесовский классификатор для многомерного случая..... 49

8.5. Обучение статистических классификаторов. Вероятность ошибок. Меры статистической разделимости..... 59

9. Самообучающиеся системы распознавания..... 63

9.1. Алгоритм корректирующих приращений..... 64

9.2. Алгоритм минимальной среднеквадратичной ошибки (НСКО-алгоритм обучения)..... 65

9.3. Метод потенциальных функций..... 66

Литература..... 68